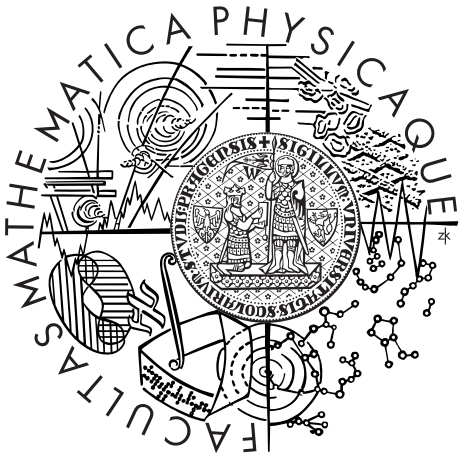
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Martin Vdovičenko

ARFIMA modely časových řad

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Šárka Hudecová, Ph.D.

Studijní program: Matematika

Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie

Praha 2013

Na tomto mieste by som sa chcel poďakovať za odborné vedenie, cenné pripomienky a konzultácie pri písaní diplomovej práce RNDr. Šárke Hudecovej, Ph.D. V neposlednom rade ďakujem za podporu mojim rodičom a priateľom.

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: ARFIMA modely časových řad

Autor: Martin Vdovičenko

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Šárka Hudecová, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Práce se zabývá procesy s dlouhou pamětí, kterou definujeme více způsoby. Hlavní pozornost je věnována modelu ARFIMA, jeho základním vlastnostem a využití. Práce dále obsahuje podrobný popis grafických, semiparametrických a parametrických metod pro odhad parametrů modelu ARFIMA. V práci uvádíme pět vybraných balíčků z programu R, které se zabývají modelováním procesů s dlouhou pamětí. Představujeme jejich základní funkce s popisem vstupních argumentů a výstupů. Na závěr aplikujeme uvedené balíčky na reálná data. Analyzujeme roční minimální výšky hladiny řeky Nil a rozebíráme výsledky dosažené různými funkcemi.

Klíčová slova: proces s dlouhou pamětí, ARFIMA model, odhady parametrů, Hurstův parametr

Title: ARFIMA time series models

Author: Martin Vdovičenko

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Šárka Hudecová, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The thesis deal with long-memory processes which are defined by several ways. The main concern is dedicated to ARFIMA model, to its basic properties and its application. Next, graphical, semiparametric and parametric estimation methods of ARFIMA parameters are described in detail. Five selected R packages are introduced that are suitable for modeling long-memory processes. We discuss their basic functions with description of input arguments and output. Finally, the application of the packages on real data is discussed according to results of each function. Data sample comes from the Nile River and represents its yearly minimal water levels.

Keywords: long-memory processes, ARFIMA model, paramater estimation, Hurst parameter

Názov práce: ARFIMA modely časových řad

Autor: Martin Vdovičenko

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci diplomovej práce: RNDr. Šárka Hudecová, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Práca sa zaoberá procesmi s dlhou pamäťou, ktorú definujeme viacerými spôsobmi. Hlavná pozornosť je venovaná modelu ARFIMA, jeho základným vlastnostiam a využitiu. Práca ďalej obsahuje podrobný popis grafických, semiparametrických a parametrických metód na odhady parametrov modelu ARFIMA. V práci uvádzame päť vybraných balíčkov z programu R, ktoré sa zaoberajú modelovaním procesov s dlhou pamäťou. Predstavujeme ich základné funkcie s popisom vstupných argumentov a výstupov. Na záver aplikujeme uvedené balíčky na reálne dáta. Analyzujeme ročné minimálne výšky hladiny rieky Níl a rozoberieme výsledky dosiahnuté rôznymi funkciami.

Kľúčové slová: proces s dlhou pamäťou, ARFIMA model, odhady parametrov, Hurstov parameter

Obsah

Úvod

Fenomén dlhej pamäte začal byť bližšie skúmaný v 50. rokoch 20. storočia v hydrologickej a klimatologickej sfére. Anglický hydrológ H. E. Hurst pri skúmaní rieky Níl zistil, že jednotlivé pozorovania sú napriek značnej časovej vzdialenosti stále významne korelované. V skutočnosti korelácie klesali v čase polynomiálne, teda pomalšie oproti exponenciálnemu poklesu pri ARMA procesoch. Tento poznatok položil základ prvotnej definície dlhej pamäte. Postupne bola vybudovaná teória s rôznymi prístupmi k definovaniu dlhej pamäte, ktoré podnietili vznik celého radu nových modelov: ARFIMA procesy, frakcionálny Brownov pohyb, frakcionálny gaussovský šum...

Neskôr našli procesy s dlhou pamäťou uplatnenie v rôznych sférach vedy. Snažili sa napríklad zodpovedať otázku zvyšujúcej sa priemernej teploty na Zemi. V 80. rokoch minulého storočia prenikli modely s dlhou pamäťou aj do makroekonomickej sféry. V ekonometrii sa dnes bežne využívaju na modelovanie cenovej volatility aktív, inflácie, indexu spotrebiteľských cien, menových kurzov, atď.

Táto práca sa podrobne zaoberá procesmi s dlhou pamäťou, pričom sa zameriava hlavne na autoregresné frakcionálne integrované procesy kĺzavých priemerov, tzv. ARFIMA procesy. Kapitola 1 ponúka základné definície a tvrdenia, ktoré sú nevyhnutné pre ďalšiu prácu. Takisto sú na tomto mieste uvedené ARMA modely so základnými vlastnosťami. V kapitole 2 zavádzame procesy s dlhou pamäťou z rôznych pohľadov a prechádzame priamo k definícii procesu ARFIMA. Táto kapitola sa ďalej zaoberá vlastnosťami ARFIMA procesov, ich predpoveďami a na záver spojením ARFIMA procesov s procesmi s dlhou pamäťou. Kapitola 3 sa zameriava na metódy odhadu parametrov modelu ARFIMA. Postupne sú podrobne uvedené grafické, semiparametrické a parametrické metódy. V kapitole 4 predstavujeme päť základných balíčkov z programu R, ktoré sa zaoberajú problematikou procesov s dlhou pamäťou. Pri jednotlivých balíčkoch uvádzame základné funkcie, ich argumenty a výstup. Na záver je prevedená analýza dát z rieky Níl využívajúca program R a predtým popísané balíčky. Práca obsahuje vlastný skript z programu R, ktorý je na priloženom disku.

Kapitola 1

Základné definície, pojmy a znalosti

1.1 Náhodný proces

Na začiatok práce zavedieme definíciu náhodného procesu, jeho základné charakteristiky a vlastnosti.

Definícia 1. *Nech* $(\Omega, \mathcal{F}, P)$ *je pravdepodobnostný priestor a nech* $T \subset R$. *Potom rodina reálnych náhodných veličín* $X = \{X_t, t \in T\}$ *definovaných na* $(\Omega, \mathcal{F}, P)$ *sa nazýva* náhodný proces.

Podľa typu množiny T rozdeľujeme náhodné procesy na procesy s diskrétnym časom ($T = \mathbb{Z}$), v tomto prípade hovoríme o časových radoch alebo náhodných postupnostiach, a na procesy so spojitým časom ($T = (a,b)$), kde (a,b) je interval na množine $\mathbb{R}$. My sa budeme ďalej zaoberať prípadom, kde $T = \mathbb{Z}$, teda časovými radmi. Ďalej zaveďme základné štatistické charakteristiky pre náhodné procesy.

Definícia 2. *Nech* $\{X_t, t \in \mathbb{Z}\}$ *je náhodný proces taký, že pre každé* $t \in \mathbb{Z}$ *existuje stredná hodnota* $\mathsf{E}\, X_t$. *Potom funkciu* $\mu_t = \mathsf{E}\, X_t$ *nazveme* stredná hodnota *procesu* $\{X_t, t \in \mathbb{Z}\}$. *V prípade, že* $\mu_t \equiv 0$, *hovoríme, že proces* $\{X_t, t \in \mathbb{Z}\}$ *je centrovaný. Nech* $\{X_t, t \in \mathbb{Z}\}$ *má navyše konečné druhé momenty, potom funkciu dvoch premenných*

$$\gamma(s,t) = cov(X_s, X_t), \quad s,t \in \mathbb{Z},$$

nazveme autokovariančná funkcia *procesu* $\{X_t, t \in \mathbb{Z}\}$. *Pod* rozptylom procesu *v čase* t, *potom rozumieme hodnotu* $\gamma(t,t)$.
Autokorelačnú funkciu procesu $\{X_t, t \in \mathbb{Z}\}$ *následne definujeme*

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)}\sqrt{\gamma(t,t)}}, \quad s,t \in \mathbb{Z}.$$

Ďalej zavedieme dôležitú vlastnosť procesov - stacionaritu.

Definícia 3. *Nech* $\{X_t, t \in \mathbb{Z}\}$ *je náhodný proces a nech pre ľubovoľné* $n \in \mathbb{N}$, *ľubovoľné* $x_1, \ldots, x_n \in \mathbb{R}$ *a* $t_1, \ldots t_n, h \in \mathbb{Z}$ *platí*

$$F_{t_1,\ldots t_n}(x_1, \ldots, x_n) = F_{t_1+h,\ldots t_n+h}(x_1, \ldots, x_n),$$

kde rovnosť v predošlej rovnici chápeme ako rovnosť pravdepodobnostných rozdelení s distribučnými funkciami $F_{t_1,\ldots t_n}$, *resp.* $F_{t_1+h,\ldots t_n+h}$. *Náhodný proces* $\{X_t, t \in \mathbb{Z}\}$ *potom nazveme striktne stacionárny.*

My však budeme pracovať s o niečo menej striktnou podmienkou stacionarity.

Definícia 4. *Náhodný proces* $\{X_t, t \in \mathbb{Z}\}$ *s konečnými druhými momentmi sa nazýva* slabo stacionárny, *ak má konštantnú strednú hodnotu* $\mu_t = \mu$ *pre všetky* $t \in \mathbb{Z}$ *a jeho autokovariančná funkcia* $\gamma(s,t)$ *je funkciou rozdielu* $s - t$.

V ďalšom texte budeme pod pojmom stacionarita uvažovať slabú stacionaritu. Stacionárny rad sa vyznačuje najmä tým, že jeho autokovariančná štruktúra nemení v čase charakter. To znamená, že u takýchto radov sa trend alebo sezónnosť nevyskytujú. Pre stacionárne procesy budeme ďalej autokovariančnú funkciu definovať ako funckiu jednej premennej $\gamma(t) = \gamma(t,0)$ a autokorelačnú funkciu ako $\rho(t) = \gamma(t)/\gamma(0)$.

Ak je každé konečnerozmerné rozdelenie náhodného procesu normálne, tak ho nazveme *gaussovský proces*. Z vlastností normálneho rozdelenia potom plynie, že pre gaussovské procesy je striktná podmienka stacionarity ekvivalentná slabej.

Podľa [17], pre každú stacionárnu náhodnú postupnosť s autokovariančnou funkciou $\gamma(t), t \in \mathbb{Z}$ platí

$$\gamma(t) = \int_{-\pi}^{\pi} e^{it\lambda} dF(\lambda), \tag{1.1}$$

kde F je sprava spojitá neklesajúca ohraničená funkcia na uzavretom intervale $[-\pi,\pi]$ a $F(-\pi) = 0$. Pravú stranu v (1.1) nazývame *spektrálny rozklad* autokovariančnej funkcie, funkciu F nazývame *spektrálna distribučná funckia* náhodnej postupnosti, pričom je vzťahom (1.1) určená jednoznačne. Ak existuje $f(\lambda) \geq 0$ pre $\lambda \in [-\pi,\pi]$ taká, že $F(\lambda) = \int_{-\pi}^{\lambda} f(x)dx$, potom sa f nazýva *spektrálna hustota*.

Na záver tejto podkapitoly ešte zavedieme vybrané vety a tvrdenia, ktoré budeme ďalej v texte používať. Začnime konvergenciou podľa kvadratického stredu pre náhodné procesy.

Definícia 5. *Hovoríme, že postupnosť náhodných veličín* $\{X_t, t \in \mathbb{Z}\}$ konverguje *k náhodnej veličine* X *podľa kvadratického stredu, ak*

$$\mathsf{E}\,|X_n - X|^2 \to 0, \quad n \to \infty.$$

Ďalej zavedieme inverzný vzorec pre spektrálnu hustotu a následne vzorec pre hustotu procesu, ktorý vznikol použitím lineárneho filtru na iný proces. Oba vzorce sú uvedené taktiež v [17], veta 3.4 a veta 5.8.

Veta 1. *Nech* $\{X_t, t \in \mathbb{Z}\}$ *je stacionárna postupnosť s absolútne sčítateľnou autokovariančnou funkciou* $\gamma(t)$*, t.j.* $\sum_{t=1}^{\infty} |\gamma(t)| < \infty$*. Potom spektrálna hustota postupnosti* $\{X_t, t \in \mathbb{Z}\}$ *existuje a je pre každé* $\lambda \in [-\pi,\pi]$ *rovná*

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} e^{-ik\lambda} \gamma(k). \tag{1.2}$$

Tvrdenie 1. *Pre proces* $X_t = \sum_{k=-\infty}^{\infty} c_k Y_{t-k}$*, ktorý vznikol použitím lineárneho fitlru* $\{c_k, k \in \mathbb{Z}\}$*, kde* $\sum_{k=-\infty}^{\infty} c_k^2 < \infty$*, na stacionárny centrovaný proces* $\{Y_t, t \in \mathbb{Z}\}$ *so spektrálnou hustotou* f_Y *a s autokovariančnou funkciou* γ_Y *platí, že je tiež stacionárny a centrovaný a pre jeho spektrálnu hustotu platí*

$$f_X(\lambda) = |\Psi(\lambda)|^2 f_Y(\lambda), \quad \lambda \in [-\pi,\pi], \tag{1.3}$$

kde $\Psi(\lambda) = \sum_{k=-\infty}^{\infty} c_k e^{-ik\lambda}$*. Pre autokovariančnú funkciu procesu* $\{X_t, t \in \mathbb{Z}\}$ *platí*

$$\gamma_X(t) = \sum_{j=-\infty}^{\infty} \sum_{k=\infty}^{\infty} c_j c_k \gamma_Y(t-j-k), \quad t \in \mathbb{Z}.$$

Poznámka 1. *Všimnime si, že oproti vete 5.8 v [17], predpokladáme o niečo slabší predpoklad* $\sum_{k=-\infty}^{\infty} c_k^2 < \infty$ *namiesto* $\sum_{k=-\infty}^{\infty} |c_k| < \infty$*. Toto zobecnenie je korektné vzhľadom k poznámke 5.2 v [17] a ako uvidíme neskôr je nevyhnutné pre triedu procesov ARFIMA.*

1.2 Dôležité modely a ich vlastnosti

V tejto podkapitole definujeme základné modely, ktoré predchádzdali vzniku samotných ARFIMA procesov. Uvedieme ich základné vlastnosti, ktoré nebudeme podrobne odvádzať, bližšie viď [4] a [17].

Definícia 6. *Náhodnú postupnosť* $\{Y_t, t \in \mathbb{Z}\}$ *nekorelovaných náhodných veličín, ktoré majú nulovú strednú hodnotu a konečný konštantý rozptyl* σ^2*, nazveme* biely šum *WN(0,*σ^2*).*

Autokovariančná funkcia bieleho šumu $\{Y_t, t \in \mathbb{Z}\} \sim \mathrm{WN}(0,\sigma^2)$ je

$$\begin{aligned} \gamma_Y(t) &= \sigma^2, & t &= 0, \\ &= 0, & t &\neq 0. \end{aligned}$$

Pre spektrálnu hustotu $\{Y_t, t \in \mathbb{Z}\}$ platí $f_Y(\lambda) = \sigma^2/(2\pi)$, kde $\lambda \in [-\pi,\pi]$.

1.2.1 Modely kĺzavých súčtov a lineárne procesy

Definícia 7. *Náhodná postupnosť* $\{X_t, t \in \mathbb{Z}\}$ *definovaná vzťahom*

$$X_t = Y_t + b_1 Y_{t-1} + \ldots + b_n Y_{t-n}, \quad t \in \mathbb{Z}, \tag{1.4}$$

kde $b_i, i = 1, \ldots, n$ *sú reálné konštanty,* $b_n \neq 0$ *a* $\{Y_t, t \in \mathbb{Z}\}$ *je biely šum WN*$(0,\sigma^2)$*, sa nazýva* postupnosť kĺzavých súčtov rádu *n. Značíme ho MA(n).*

Proces MA(n) je centrovaný a stacionárny. Jeho autokovariančná funkcia sa počíta priamo z definície kovariancie a platí pre ňu

$$\begin{aligned} \gamma_X(t) &= \sigma^2 \sum_{k=0}^{n-|t|} b_{k+|t|} b_k, && |t| \leq n, \\ &= 0, && |t| > n, \end{aligned}$$

kde $b_0 = 1$. Bod useknutia k_0 autokorelačnej funkcie, t.j. čas, od ktorého je autokorelačná funkcia nulová, je prirodzene vzhľadom k vzorcu pre autokovariančnú funkciu $k_0 = n$.

Spektrálna hustota procesu MA(n) existuje a platí pre ňu priamo z tvrdenia 1 nasledujúci vzťah

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} \left| \sum_{k=0}^{n} b_k e^{-ik\lambda} \right|^2, \quad \lambda \in [-\pi, \pi].$$

Zobecnením modelu MA(n) môžeme prejsť k definícii lineárneho procesu.

Definícia 8. *Nech* $\{Y_t, t \in \mathbb{Z}\}$ *je biely šum WN*$(0,\sigma^2)$ *a nech je postupnosť konštánt* $\{c_j, j \in \mathbb{N}_0\}$ *taká, že* $\sum_{j=0}^{\infty} |c_j| < \infty$*. Potom postupnosť definovaná predpisom*

$$X_t = \sum_{j=0}^{\infty} c_j Y_{t-j}, \quad t \in \mathbb{Z}, \tag{1.5}$$

sa nazýva kauzálny lineárny proces*, značíme* MA(∞)*.*

Nekonečný súčet z predošlej definície chápeme v tom zmysle, že X_t je limitou podľa kvadratického stredu veličín $\sum_{j=0}^{n} c_j Y_{t-j}$ pre každé t a n idúce do nekonečna. Existenciu tejto limity nám zaručuje predpoklad na absolútnu sčítateľnosť konštánt c_j, viď napr. [5] alebo [17].

Kauzalita v definícii 8 znamená, že náhodná veličina X_t je vyjadrená len pomocu minulých a súčasnej veličiny $Y_s, s \leq t$. Kauzalita zohráva dôležitú úlohu najmä pri konštrukcii predpovedí. Proces MA(∞) je centrovaný a stacionárny. Pre jeho autokovariančnú funkciu platí

$$\gamma_X(t) = \sigma^2 \sum_{k=0}^{\infty} c_{k+|t|} c_k.$$

Spektrálna hustota procesu *MA(n)* existuje a platí pre ňu

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} \left| \sum_{k=0}^{\infty} c_k e^{-ik\lambda} \right|^2, \quad \lambda \in [-\pi, \pi].$$

Poznámka 2. *V ďalšom texte budeme pod pojmom jednotkový kruh rozumieť množinu* $\{z : |z| \leq 1, z \in \mathbb{C}\}$. *Jednotkovú kružnicu budeme definovať ako* $\{z : |z| = 1, z \in \mathbb{C}\}$.

Definícia 9. *Lineárny proces* $\{X_t, t \in \mathbb{Z}\}$ *sa nazýva* invertibilný, *ak exsituje postupnosť konštánt* $\{e_j, j \in \mathbb{N}_0\}$ *taká, že* $\sum_{j=0}^{\infty} |e_j| < \infty$ *a platí*

$$Y_t = \sum_{j=0}^{\infty} e_j X_{t-j}, \quad t \in \mathbb{Z}, \tag{1.6}$$

kde $\{Y_t, t \in \mathbb{Z}\}$ *je biely šum WN*$(0,\sigma^2)$.

Postačujúcou podmienkou na to, aby bol lineárny proces invertibilný je, aby všetky korene polynómu $b(z) = 1 + b_1 z + \ldots + b_n z^n$ ležali mimo jednotkového kruhu.

1.2.2 Autoregresné modely

Definícia 10. *Náhodná postupnosť* $\{X_t, t \in \mathbb{Z}\}$ *sa nazýva* autoregresná postupnosť rádu *m (označujeme AR(m)), ak spĺňa*

$$X_t + a_1 X_{t-1} + \ldots + a_m X_{t-m} = Y_t, \quad t \in \mathbb{Z}, \tag{1.7}$$

kde $a_i, i = 1, \ldots m$ *sú reálné konštanty,* $a_m \neq 0$ *a* $\{Y_t, t \in \mathbb{Z}\}$ *je biely šum WN*$(0,\sigma^2)$.

Postačujúcou podmienkou na to, aby sa proces AR(m) dal previesť na kauzálny lineárny proces, je to, aby polynóm $a(z) = 1 + a_1 z + \ldots + a_m z^m$ mal všetky korene mimo jednotkového kruhu. V tomto prípade je zmienená podmienka zároveň postačujúcou podmienkou pre stacionaritu. Pre koeficienty c_j z vyjdrenia (1.5) platí

$$c(z) = \sum_{j=0}^{\infty} c_j z^j = \frac{1}{a(z)}, \quad |z| \leq 1.$$

Autokovariančnú funkciu, pre ktorú neexistuje bod useknutia, a spektrálnu hustotu potom odvodíme analogicky ako pri lineárnom procese. Autokovariančnú funkciu autoregresného procesu, ktorý sa dá vyjadriť ako kauzálny lineárny proces, môžeme vypočítať aj pomocou tzv. *Yuleových-Walkerových rovníc*. Táto metóda využíva nekorelovanosť bieleho šumu Y_t z rovnice (1.7) s náhodnými veličinami X_s, kde $s < t$, ktorá plynie z kauzality $\{X_t, t \in \mathbb{Z}\}$. Postupne násobíme obe

strany rovnice veličinami Y_t, X_{t-k} a aplikujeme strednú hodnotu. Takto získavame sústavu diferenčných rovníc pre parametre a_i, $i = 1, \ldots, m$

$$\begin{aligned} \gamma(0) + a_1\gamma(1) + \ldots + a_n\gamma(n) &= \sigma^2, & k &= 0 \\ \gamma(k) + a_1\gamma(k-1) + \ldots + a_n\gamma(n-k) &= 0, & k &\geq 1. \end{aligned}$$

Pod označením AR(∞) budeme rozumieť model

$$Y_t = \sum_{j=0}^{\infty} a_j X_{t-j}, \quad t \in \mathbb{Z},$$

kde $\{Y_t, t \in \mathbb{Z}\}$ je biely šum WN(0,σ^2).

1.2.3 Zmiešaný model ARMA

Definícia 11. *Náhodná postupnosť $\{X_t, t \in \mathbb{Z}\}$ sa riadi modelom ARMA(m,n), ak*

$$X_t + a_1X_{t-1} + \ldots + a_mX_{t-m} = Y_t + b_1Y_{t-1} + \ldots + b_nY_{t-n}, \quad t \in \mathbb{Z},$$

kde $a_i, i = 1, \ldots m, b_j, j = 1, \ldots, n$, sú reálné konštanty, $a_m \neq 0, b_n \neq 0$ a $\{Y_t, t \in \mathbb{Z}\}$ je biely šum WN(0,σ^2). Model ARMA(m,n) nazývame zmiešaný proces.

Na tomto mieste je vhodné definovať operátor posunutia B, ktorý je praktický pri procesoch tohto typu:

$$BX_t = X_{t-1},$$
$$B^kX_t = B^{k-1}(BX_t) = X_{t-k}, \quad k \in \mathbb{Z}.$$

Symbolu B^kX_t hovoríme k-ta diferencia radu. Model ARMA(m,n) teraz vieme pomocou operátoru posunutia napísať v tvare

$$a(B)X_t = b(B)Y_t. \tag{1.8}$$

Operátor $a(B) = 1 + a_1B + \ldots + a_mB^m$ nazveme *autoregresný operátor* a naopak $b(B) = 1 + b_1B + \ldots + b_nB^n$ nazveme *operátor kĺzavých súčtov*. Pričom vidíme, že operátory $a(B)$, resp. $b(B)$ sú formálne zhodné s algebraickým polynómom rádu m, resp. n. Model ARMA(m,n) sa nazýva zmiešaný model autoregresie a kĺzavých súčtov, pretože autoregresný model rádu m AR(m) a model kĺzavých súčtov rádu n MA(n) sú jeho špeciálnymi prípadmi.

Ak polynómy $a(z)$ a $b(z)$ nemajú žiadne spoločné korene a nech polynóm $a(z)$ má všetky korene mimo jednotkového kruhu, tak potom sa dá proces ARMA(m,n) previesť na proces MA(∞), pričom pre koeficienty c_j z (1.5) platí

$$c(z) = \sum_{j=0}^{\infty} c_j z^j = \frac{b(z)}{a(z)}, \quad |z| \leq 1.$$

Za týchto podmienok je proces ARMA(m,n) stacionárny a pre jeho spektrálnu hustotu platí

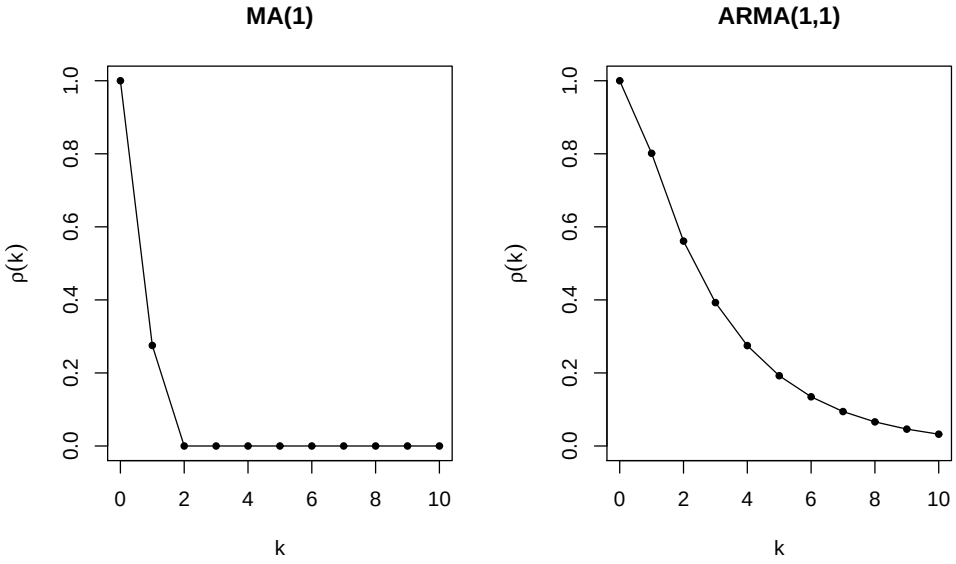
$$f_X(\lambda) = \frac{\sigma^2}{2\pi} \frac{|\sum_{j=0}^{n} b_j e^{-ij\lambda}|^2}{|\sum_{k=0}^{m} a_k e^{-ik\lambda}|^2}, \quad \lambda \in [-\pi, \pi], \tag{1.9}$$

kde $a_0 = 1, b_0 = 1$.

Podmienka na prevod procesu ARMA(m,n) na tvar (1.6), t.j. podmienka invertibility, je totožná s podmienkou na proces MA(∞) s tým, že tentokrát sa požaduje, aby korene polynómu $b(z)$ ležali mimo jednotkového kruhu. V tomto prípade pre koeficienty e_j z (1.6) platí

$$e(z) = \sum_{j=0}^{\infty} e_j z^j = \frac{a(z)}{b(z)}, \quad |z| \leq 1.$$

Dôkazy a presné znenia viet sú dostupné v [17].



Obr. 1.1: Autokorelačná funkcia procesov MA(1) a ARMA(1,1)

Všimnime si, že aj pre zmiešaný model ARMA(m,n), ktorý sa dá vyjadriť ako lineárny proces, môžeme autokovariančnú, resp. autokorelačnú funkciu vypočítať z analogickej sústavy diferenčných rovníc, ako pri autoregresnom modeli. Sústava má tvar

$$\gamma(k) + a_1\gamma(k-1) + \ldots + a_m\gamma(k-m) = \sigma^2 \sum_{j=k}^{n} b_j c_{j-k}, \qquad 0 \leq k \leq n,$$

$$\gamma(k) + a_1\gamma(k-1) + \ldots + a_m\gamma(k-m) = 0, \qquad k > n,$$

kde sme využili kauzalitu lineárneho procesu a tým pádom nekorelovanosť X_{t-j}, Y_{t-k} pre $k < j$. Báza riešení takýchto diferenčných rovníc je tvorená exponenciálami, t.j. autokorelačná funkcia procesu ARMA (AR) nemá bod useknutia.

V skutočnosti je lineárnou kombináciou klesajúcich geometrických postupností a sínusoid rôznych frekvencií s klesajúcimi amplitúdami s výnimkou prvých $n-m$ hodnôt, ak $n \geq m$, viď [5]. To znamená, že autokorelačná funckcia ARMA procesu klesá k nule exponenciálne a je ohraničená

$$|\rho(k)| \leq Cr^{-k},$$

pre nejaké $C > 0$ a $0 < r < 1$.

Na obrázku 1.1 sme si nechali vykresliť autokorelačnú funkciu procesov MA(1) s $b_1 = 0{,}3$ a ARMA(1,1) s $a_1 = -0{,}7$ a $b_1 = 0{,}3$. U MA(1) procesu pozorujeme bod uskenutia v čase 1. U ARMA procesu zase môžeme pozorovať exponenciálny pokles k 0. Autokovariančná funkcia procesu AR(1) je zobrazená na obrázku 2.4.

Poznámka 3. *Doteraz sme uvažovali centrované procesy. Pre procesy s nenulovou a v čase nemennou strednou hodnotou μ by sme však postupovali analogicky. Napríklad proces ARMA(m,n) by mal tvar*

$$(X_t - \mu) + a_1(X_{t-1} - \mu) + \ldots + a_m(X_{t-m} - \mu) = Y_t + b_1 Y_{t-1} + \ldots + b_n Y_{t-n}$$

Kapitola 2

Náhodné procesy s dlhou pamäťou

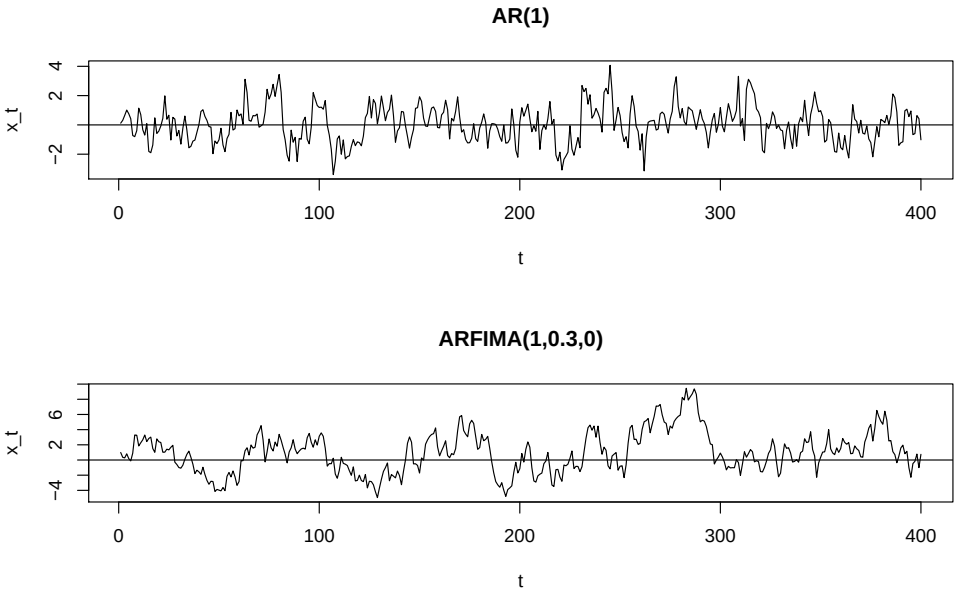
2.1 Definícia dlhej pamäte

Počiatky skúmania procesov s dlhou pamäťou siahajú do začiatku druhej polovice 20. storočia. V rôznych sférach výskumu (hydrológia, geológia, astronómia, atď.) prichádzali vedci do styku s dátami, ktoré sa nedali podchytiť dovtedy známymi modelmi. Tieto dáta boli charakteristické hlavne tým, že korelácie medzi jednotlivými pozorovaniami klesali v čase k nule pomalšie ako u dovtedy používaných modelov. Pre tieto dáta už nepostačovali populárne modely typu ARMA. Vznikla preto nová skupina procesov, ktorej hovoríme procesy s dlhou pamäťou. Tieto procesy sa vyznačujú hlavne tým, že

- dáta obsahujú periódy, v ktorých pozorovania majú buď veľké alebo malé hodnoty,
- skúmaním len malej časovej vzorky dát pozorujeme trendy alebo cykly, avšak pozorovaním celého vzorku žiadny trend alebo opakujúce sa cykly nepozorujeme,
- korelácie v čase klesajú k nule polynomiálne,
- rozptyl výberového priemeru klesá k nule pomalšie ako n^{-1}, a to proporčne k $n^{-\alpha}$, kde $\alpha \in (0,1)$,
- spektrálna hustota má pól v nule.

Medzi najznámejšie procesy s dlhou pamäťou patria ARFIMA procesy, frakcionálny Brownov pohyb a frakcionálny gaussovský šum. Pre názornú ukážku sú na obrázku 2.1 zobrazené priebehy procesu AR(1) s $a_1 = -0{,}7$ a procesu ARFIMA$(1; 0.3; 0)$ s $a_1 = -0{,}7$. Proces AR(1) je proces s krátkou pamäťou,

na druhej strane ARFIMA$(1;0{,}3;0)$ je proces s dlhou pamäťou. Na prvý pohľad vidíme, že detekovať proces s dlhou pamäťou sa z realizácie procesu dá len veľmi obtiažne. Vidíme, že u ARFIMA procesu existujú cykly, ktoré sú však nepravidelné. Takisto vidíme, že na rozdiel od modelu AR, kde hodnoty prudko „skáču“ okolo nuly, hodnoty procesu s dlhou pamäťou majú tendenciu zotrvať na určitej úrovni.



Obr. 2.1: Priebeh procesov AR(1) a ARFIMA(1;0,3;0)

Možností ako definovať proces s dlhou pamäťou je viacero.

Definícia 12. *Nech* $\{X_t, t \in \mathbb{Z}\}$ *je stacionárny proces a* $\rho(h)$ *jeho autokorelačná funkcia. Nech existuje reálne číslo* $\alpha \in (0{,}1)$ *a konštanta* $c_\rho > 0$ *také, že*

$$\lim_{k\to\infty} \frac{\rho(k)}{c_\rho k^{-\alpha}} = 1. \tag{2.1}$$

Potom hovoríme, že stacionárny proces $\{X_t, t \in \mathbb{Z}\}$ *je* proces s dlhou pamäťou.

Táto definícia popisuje len asymptotickú rýchlosť klesania korelácií k 0, nehovorí nám nič o samotných koreláciach $\rho(k)$. Z definície 12 vyplýva, že $\rho(k) \sim c_\rho(k)|k|^{-\alpha}$ a keďže $\alpha \in (0{,}1)$, tak platí

$$\sum_{k=-\infty}^{\infty} |\rho(k)| = \infty. \tag{2.2}$$

Teda korelácie medzi jednotlivými pozorovaniami neklesajú k 0 dostatočne rýchlo (klesajú polynomiálne) na to, aby suma korelácií bola sčítateľná. Procesy so sčítateľnými koreláciami nazývame *procesy s krátkou pamäťou*. Typickým príkladom

procesov s krátkou pamäťou sú ARMA procesy. Sčítateľnosť korelácií (resp. kovariancií) je predpokladom viacerých teoretických viet, napr. postačujúca podmienka pre existenciu spektrálnej hustoty, asymptotické rozdelenie výberového priemeru, atď. Viac viď. [17].

Namiesto parametru α sa v praxi často používa parameter $H = 1 - \frac{\alpha}{2}$. Parameter H je v literatúre označovaný ako *parameter dlhej pamäte* alebo *Hurstov parameter*. Parameter H nadobúda hodnoty v intervale (0,1). Hodnotu 0,5 interpretujeme tak, že proces nemá dlhú pamäť, t.j má krátku pamäť. Čím sa H blíži viac k 1, tým sa v rade prejavuje dlhá pamäť silnejšie. Pre H menšie ako 0,5 hovoríme, že má proces strednú pamäť. Konkrétnejšie vysvetlenie týchto tvrdení je poskytnuté v časti o procese ARFIMA.

Ak poznáme korelačnú, resp. kovariančnú funkciu, tak k definícii procesu s dlhou pamäťou sa dá pristúpiť ekvivalentne aj pomocou spektrálnej hustoty.

Definícia 13. *Nech X_t je stacionárny proces a nech existuje reálne číslo $\beta \in (0,1)$ a konštanta $c_f > 0$ také, že*

$$\lim_{\lambda \to 0} \frac{f(\lambda)}{c_f |\lambda|^{-\beta}} = 1. \tag{2.3}$$

Stacionárny proces $\{X_t, t \in \mathbb{Z}\}$ potom nazveme proces s dlhou pamäťou.

Z tejto definície vidíme, že spektrálna hustota má v nule naozaj pól, tzn. jej limita sprava v bode nula prekračuje všetky medze. Tieto dve definície sú ekvivalentné v tom zmysle, že pre parametre α a β platí

$$H = 1 - \frac{\alpha}{2}, \quad H = \frac{\beta}{2} + \frac{1}{2}, \tag{2.4}$$

viď. napr. [2].

V literatúre sa však môžme stretnúť s obecnejšou definíciou procesov s dlhou pamäťou. V nich vystupujú namiesto konštánt c_ρ a c_f z definícií 12 a 13 tzv. *pomaly meniace sa funkcie*, pre ktoré platí $L(tx)/L(x) \to 1, t \in \mathbb{R}, x \to \infty$. Medzi pomaly sa meniace funkcie patria všetky spojité funkcie a teda aj konštanty. Pre ďalší vývoj práce sa nám stačí v definíciach obmedziť práve na konštanty.

Sebepodobné procesy

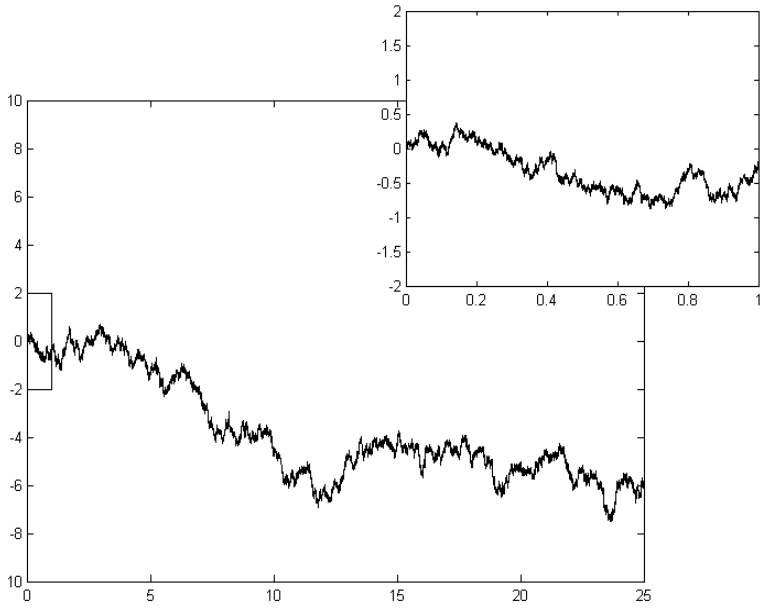
Existuje aj obecnejší prístup k procesom s dlhou pamäťou, ktorý vychádza z triedy procesov, ktoré majú vlastnosť *self-similarity*, nazývame ich tiež *sebepodobné procesy*. Pri tomto prístupe napríklad nemusíme predpokladať existenciu autokovariančnej funkcie. Uveďme definíciu self-similary procesu.

Definícia 14. *Nech $H > 0$. Náhodný proces $\{X_t, t \in \mathbb{Z}\}$ nazývame H-sebepodobný, ak platí*

$$F(c^H X_{t_1}, c^H X_{t_2}, \ldots, c^H X_{t_n}) = F(X_{ct_1}, X_{ct_2}, \ldots, X_{ct_n})$$

pre všetky $c > 0$, $n \in N$ a $t_i \geq 0$, kde $i = 1, \ldots, n$.

Symbol F značí konečne rozmerné rozdelenie a rovnosť v definícii je rovnosťou rozdelení náhodných procesov. Trajektórie self-similar procesu pôsobia tak, že majú rovnaký priebeh nezávisle na tom, z akej „diaľky“ sa na ne pozeráme, viď obrázok 2.2. Objekty s touto vlastnosťou sa označujú súhrnným pomenovaním *fraktál*.



Obr. 2.2: Vlastnosť self-similarity

Ak má sebepodobný proces $\{X_t, t \in \mathbb{Z}\}$ stacionárne prírastky $U_t = X_t - X_{t-1}$, tak pre autokorelačnú funkciu procesu U_t platí vzorec

$$\rho(k) = \frac{1}{2}[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}], \tag{2.5}$$

pre $k \geq 0$, pričom platí $\rho(k) = \rho(-k)$, pre $k < 0$, viď [2]. Pre funkciu ρ, môžeme písať $\rho(k) = \frac{1}{2}k^{2H}g(k^{-1})$, kde $g(x) = (1+x)^{2H} - 2 + (1-x)^{2H}$. Ak funkciu g rozvinieme do Taylorovho radu v nule pre $H \in (0,1)\backslash\{\frac{1}{2}\}$, tak dostávame prvý nenulový člen $2H(2H-1)x^2$. Potom z (2.5) máme

$$\frac{\rho(k)}{H(2H-1)k^{2H-2}} \to 1, \quad k \to \infty. \tag{2.6}$$

Z (2.6) je zrejmé, že proces U_t má pre $0{,}5 < H < 1$ dlhú pamäť v zmysle definície 2.1. Pre $0 < H < 0{,}5$ má proces strednú pamäť a pre $H = 0{,}5$ sú všetky korelácie procesu nulové, proces má krátku pamäť.

Ukázali sme, že ak je daný sebepodobný proces so stacionarnánymi prírastkami a Hurstovým koeficientom $0{,}5 < H < 1$, tak rad jeho diferencií je stacionárny rad s dlhou pamäťou. Na druhej strane vieme ukázať, že pre daný stacionárny proces s dlhou pamäťou platí, že jeho čiastočné súčty sú po vhodnom preškálovaní sebepodobný proces s $0{,}5 < H < 1$.

Poznámka 4. *Všimnime si, že doteraz sme uvažovali stále procesy s konečnými druhými momentmi. Pre procesy, ktoré túto vlastnosť nemajú, nemusia hore uvedené poznatky platiť. Bližsie sa problematikou zaoberá Samoridnitsky a Taqqu (1993).*

2.2 Model ARFIMA

V praxi sa stretávame s časovými radmi, ktoré nie sú stacionárne. Pri určitom type nestacionarity sa dá pristúpiť k vhodnej diferencii pôvodného radu a následne modelovať nový stacionárny rad, ktorý vznikol diferencovaním toho pôvodného.

Ak vzťah (1.8) platí pre d-tu diferenciu procesu $\{X_t, t \in \mathbb{Z}\}$, kde d je prirodzené číslo, definujeme novú skupinu náhodných procesov, tzv. *integrovaný zmiešaný proces rádu m,d,n ARIMA(m,d,n)*:

$$a(B)(1-B)^d X_t = b(B)Y_t. \tag{2.7}$$

Poznamenajme ešte, že ak platí (2.7) pre $d \geq 1$, tak pôvodný rad nie je stacionárny. Stacionárna je až jeho d-ta diferencia. V skutočnosti modelujeme novovzniknutý rad $(1-B)^d X_t$ modelom ARMA(m,n). Pripomeňme, že postačujúcou podmienkou stacionarity pre proces riadiaci sa modelom ARMA(m,n) je, že všetky korene autoregresného polynómu $a(z)$ ležia mimo jednotkového kruhu komplexnej roviny.

Model (2.7) dokážeme zobecniť, ak pripustíme, aby d mohlo byť reálne. Najprv pre $d \geq 0, d \in \mathbb{Z}$ platí známe

$$(1-B)^d = \sum_{k=0}^{d} \binom{d}{k} (-1)^k B^k. \tag{2.8}$$

Binomické koeficienty v predošlom vzťahu môžeme vyjadriť pomocou gama funkcie Γ

$$\binom{d}{k} = \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}. \tag{2.9}$$

Funkciu Γ definujeme nasledovne

$$\begin{aligned}\Gamma(x) &= \int_0^\infty t^{x-1}e^{-t}dt, && x > 0,\\ &= \infty, && x = 0,\\ &= x^{-1}\Gamma(1+x), && x < 0.\end{aligned}$$

Keďže Γ funkcia je konečná na celej množine všetkých reálnych čísel okrem nuly a záporných celých čísel, môžeme (2.8) rozšíriť na obecný prípad, kedy $d \in R\backslash\{0,-1,-2,...\}$

$$(1-B)^d = \sum_{k=0}^{\infty}\binom{d}{k}(-1)^k B^k. \tag{2.10}$$

Ako uvidíme neskôr, prípadom, kedy je d záporné celé, sa zaoberať nebudeme. Ak d je nula, tak sa jedná o ARMA proces. Nakoniec môžeme pristúpiť k definícii samotného modelu ARFIMA.

Definícia 15. *Nech $\{X_t, t \in \mathbb{Z}\}$ je stacionárny proces a nech existuje $d \in (-\frac{1}{2}, \frac{1}{2})$ také, že*

$$a(B)(1-B)^d X_t = b(B)Y_t. \tag{2.11}$$

Potom $\{X_t, t \in \mathbb{Z}\}$ nazývame frakcionálne integrovaný zmiešaný proces rádu *m, d, n, značený* ARFIMA*(m,d,n).*

Táto definícia bola zavedená v [11]. Podstatným z definície 15 je obmedzenie parametru d na interval $(-0{,}5; 0{,}5)$. Je to z dôvodu, že práve na tomto intervale je proces (za vhodných podmienok pre polynómy $a(z)$ a $b(z)$, viď. ďalej) stacionárny, invertibilný a kauzálny.

ARFIMA proces, pre ktorý $m = 0$ a $n = 0$, sa niekedy nazýva frakcionálny (čiastočný) šum a označuje sa FI(d).

Pre $d \geq 0{,}5$ je proces nestacionárny, naopak pre $d \leq -0{,}5$ je proces neinvertibilný. Zároveň platí, že, ak je $d \geq 0{,}5$, môžeme vhodnou diferenciou proces upraviť na prípad, kedy $d \in (-0{,}5; 0{,}5)$, viď. [11].

Ak $d \in (0; 0{,}5)$, tak pre autokorelačnú funkciu platí

$$\sum_{k=-\infty}^{\infty}\rho(k) = \infty.$$

Toto tvrdenie plynie priamo z inverzného vzorcu (1.2) pre spektrálnu hustotu stacionárnej postupnosti s absolútne sčítateľnou autokovariančnou funkciou. Ak pre spor predpokladáme, že pre $d \in (0; 0{,}5)$ je suma autokovariancií konečná, tak môžeme použiť spomínaný inverzný vzorec. Ako uvidíme neskôr, viď podkapitola

2.2.1, v nule má spektrálna hustota ARFIMA procesu pól, teda dostávame spor pre sčítateľnosť autokovariancií. V tomto prípade, viď (2.2), je proces riadiaci sa modelom ARFIMA proces s dlhou pamäťou.

Naopak pre $d \in (-0{,}5; 0)$ platí

$$\sum_{k=1}^{\infty} |\rho(k)| < \infty.$$

Pre tento prípad používame pomenovanie proces so strednou pamäťou, viď. [5]. Toto tvrdenie plynie priamo z tvrdenia 1.

V [16] sa používa v definícii ARFIMA modelu menej obmedzujúca podmienka na parameter d a pripúšťajú sa hodnoty $d < 0{,}5$ a $d \neq 0, -1, -2, \ldots$, t.j. mimo pólov funkcie Γ. Ďalej sa však už pracuje s intervalom $d \in (-1; 0{,}5)$.

Predchádzajúce tvrdenia môžeme sformulovať do vety.

Veta 2. *Nech $\{X_t, t \in \mathbb{Z}\}$ je proces riadiaci sa ARFIMA modelom v zmysle definície 15. Nech polynómy $a(z), b(z)$ nemajú žiadne spoločné korene. Potom platí:*

1. *Ak korene autoregresného polynómu $a(z)$ ležia mimo jednotkovej kružnice, tak $\{X_t, t \in \mathbb{Z}\}$ je stacionárny a je rovnicou (2.11) určený jednoznačne, pričom platí*

$$X_t = \sum_{j=-\infty}^{\infty} c_j Y_{t-j}, \tag{2.12}$$

   kde $c(z) = (1-z)^{-d} \frac{b(z)}{a(z)}$.

2. *Ak korene autoregresného polynómu $a(z)$ ležia mimo jednotkového kruhu, tak $\{X_t, t \in \mathbb{Z}\}$ je kauzálny.*

3. *Ak korene polynómu kĺzavých súčtov $b(z)$ ležia mimo jednotkového kruhu, tak $\{X_t, t \in \mathbb{Z}\}$ je invertibilný.*

Dôkaz. Dôkaz *2* a *3* je takmer totožný dôkazu obdobných vlastností pre ARMA modely.

Pri dokazovaní stacionarity najprv ukážeme, že frakcionálny šum $Z_t = (1 - B)^{-d} Y_t = \sum_{j=0}^{\infty} \psi_j Y_{t-j}$ je dobre definovaný stacionárny proces. Neskôr bude uvedené, aký majú ψ_j asymptotický tvar, viď rovnica (2.22). Z nej vyplýva, že $\sum_{j=1}^{\infty} \psi_j^2 < \infty$, pretože uvažujeme proces s dlhou pamäťou, teda $d \in (-0{,}5; 0{,}5)$. Podľa vety 5.2 (1) v [17] je Z_t dobre definovaná v zmysle konvergencie podľa stredu. Následným použitím tvrdenia 1 získavame stacionaritu.

Dôkaz ďalej pokračuje ukázaním platnosti vzťahu $c(z) = (1-z)^{-d}\frac{b(z)}{a(z)}$ a štandardným dôkazom jednoznačnosti. Podrobne viď [4] a [16].

❑

Na rovnicu (2.11) sa môžeme pozerať z viacerých pohľadov. Na jednej strane, ak na $\{X_t, t \in \mathbb{Z}\}$ aplikujeme diferenčný operátor $(1-B)^d$, tak dostaneme ARMA proces $\tilde{X}_t$

$$\tilde{X}_t = (1-B)^d X_t. \tag{2.13}$$

Na druhej strane môžeme písať

$$X_t = a(B)^{-1} b(B) X_t^*,$$

kde X_t^* je ARFIMA(0,d,0) proces

$$X_t^* = (1-B)^{-d} Y_t.$$

Parameter d definuje proces z hľadiska dlhodobého vývoja, naopak parametre a_i a b_j, $i = 1, \ldots, m$,$j = 1, \ldots, n$ modelujú krátkodobý priebeh.

2.2.1 Spektrálna hustota

Využijeme ARMA proces $\tilde{X}_t$ definovaný vzťahom (2.13). Vieme, že pre spektrálnu hustotu procesu ARMA(m,n) platí vzorec (1.9), ktorý môžeme prepísať do pohodlnejšieho tvaru

$$\tilde{f}(\lambda) = \frac{\sigma_Y^2 |b(e^{i\lambda})|^2}{2\pi |a(e^{i\lambda})|^2}. \tag{2.14}$$

Potom, ak proces X_t vznikol z $\tilde{X}_t$ aplikovaním operátoru $(1-B)^{-d}$, tak pre spektrálnu hustotu ARFIMA procesu $\{X_t, t \in \mathbb{Z}\}$ platí podľa (1.3) z tvrdenia 1

$$f(\lambda) = |1 - e^{i\lambda}|^{-2d} \tilde{f}(\lambda) = \frac{\sigma_Y^2}{2\pi} \left(2 \sin \frac{\lambda}{2}\right)^{-2d} \frac{|b(e^{i\lambda})|^2}{|a(e^{i\lambda})|^2}. \tag{2.15}$$

Keďže $\lim_{\lambda \to 0} \lambda^{-1}(2 \sin \frac{\lambda}{2}) = 1$, tak v nule sa spektrálna hustota správa nasledovne

$$f(\lambda) \sim \tilde{f}(0) |\lambda|^{-2d}.$$

Z toho plynú dva fakty: pre $d > 0$ má spektrálna hustota v nule pól a ako sme ukázali hustota (2.15) procesu ARFIMA pre $d \in (0, \frac{1}{2})$ spĺňa podmienku dlhej pamäte (2.3). Z (2.4) plynie

$$d = H - \frac{1}{2}. \tag{2.16}$$

Na obrázku 2.3 je zobrazená hustota ARFIMA procesu z obrázku 2.1, na ktorom je zreteľne vidieť spomínaný pól v nule.

Na odhad spektrálnej hustoty sa v praxi používa tzv. *periodogram* definovaný ako

$$I(\lambda) = \frac{1}{2\pi n}\left|\sum_{j=1}^{n} X_j e^{i\lambda j}\right|^2 .$$

Pre stacionárne rady s ohraničenou spektrálnou hustotou platí

$$I(\lambda) \to_d f(\lambda)\eta, \quad n \to \infty$$

kde η je exponenciálna náhodná veličina so strednou hodnotou 1 a symbol $\to_d$ označuje konvergenciu v distribúcii. Uvažujme ďalej frekvencie $\lambda_{n,1}, \ldots, \lambda_{n,k}$, pričom

$$\lambda_{n,j} \to \lambda_j, \quad n \to \infty$$

kde $\lambda_j \pm \lambda_{j'} \neq 2\pi l$, pre $j \neq j'$. Potom tiež platí

$$[I(\lambda_{n,1}), \ldots, I(\lambda_{n,k})] \to_d [f(\lambda_1)\eta_1, \ldots, f(\lambda_k)\eta_k], \tag{2.17}$$

kde $\eta_1, \ldots, \eta_k$ sú znova exponenciálne náhodné veličiny so strednou hodnotou 1. Toto tvrdenie platí aj pre náhodné procesy s dlhou pamäťou, pre ktoré existuje vyjadrenie MA(∞), viď [2] (veta 3.7). Periodogram je teda asymptoticky nestranným odhadom spektrálnej hustoty pre $\lambda \neq 0$ a platí

$$\lim_{n\to\infty} \mathsf{E}\,[I(\lambda)] = f(\lambda).$$

2.2.2 Autokovariančná a autokorelačná funkcia

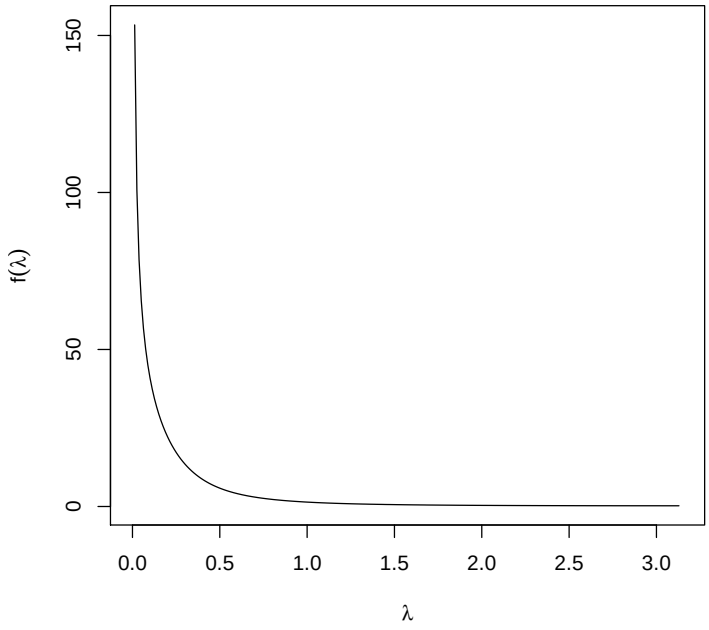
Začnime najprv modelom ARFIMA(0,*d*,0). Pre tento model nie je výpočet autokovariančnej a autokorelačnej funkcie náročný. Odvodenie vzťahov nižšie, viď [10]. Pre autokovariančnú funkciu platí

$$\begin{aligned}\gamma_0(h) &= \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda) d\lambda \\ &= \frac{\sigma_Y^2}{\pi}\int_0^{\pi} \cos(h\lambda)(2\sin(\lambda/2))^{-2d} d\lambda \\ &= \sigma_Y^2 \frac{\Gamma(1-2d)}{\Gamma(1-d)\Gamma(d)}\frac{\Gamma(h+d)}{\Gamma(1+h-d)},\end{aligned} \tag{2.18}$$

pričom sme využili rovnicu hustoty (2.15) pre ARFIMA(*0,d,0*) a identitu

$$\int_0^{\pi} \cos(hx)\sin^{v-1}(x)dx = \frac{\pi\cos(h\pi/2)\Gamma(v+1)2^{1-v}}{v\Gamma((v+h+1)/2)\Gamma((v-h+1)/2)}.$$

Pre autokorelačnú funkciu potom priamo z (2.18) platí



Obr. 2.3: Spektrálna hustota procesu ARFIMA(1,0.3,0)

$$\rho_0(h) = \frac{\Gamma(1-d)}{\Gamma(d)} \frac{\Gamma(h+d)}{\Gamma(1+h-d)}.$$

Aplikáciou Stirlingovej formuly

$$\Gamma(x) \sim \sqrt{2\pi} e^{-x+1} (x-1)^{x-\frac{1}{2}}, \quad x \to \infty, \tag{2.19}$$

na autokorelačnú funkciu ρ_0 modelu ARFIMA(*0,d,0*) dostávame asymptotický vývoj, kde

$$\rho_0(h) \sim h^{2d-1} \frac{\Gamma(1-d)}{\Gamma(d)}, \quad h \to \infty. \tag{2.20}$$

Týmto sme tiež ukázali, že frakcionálny šum je proces s dlhou pamäťou v zmysle definície 12 a zároveň jeho korelácie nie sú sčítateľné.

Vzhľadom k zložitosti zápisu autokovariančnej, resp. autokorelačnej funkcie obecného procesu ARFIMA(m,d,n) ju v tejto práci nebudeme uvádzať. Môžeme uviesť aspoň jej asymptotické správanie. Presné vzorce viď [10]. Dá sa ukázať, že pre autokovariančnú funkciu platí

$$\gamma(h) \sim c_\gamma |h|^{2d-1}, \quad |h| \to \infty, \tag{2.21}$$

kde

$$c_\gamma = \frac{\sigma^2}{\pi}\frac{|b(1)|^2}{|a(1)|^2}\Gamma(1-2d)\sin(\pi d).$$

Pri odvodzovaní vyššie uvedených vzťahov sa postupuje tak, že autokovariančnú funkciu obecného procesu vyjadríme ako konvolúciu autokovariančných funkcií (viď tiež (3.12)) časti ARMA(m,n), označme ju γ_1, a frakcionálneho šumu FI(d), označme ju γ_0. Túto operáciu nám priamo umožňuje použiť tvrdenie 1. Potom platí

$$\gamma(h) = \sum_{k=0}^{\infty} \gamma_1(k)\gamma_0(h-k).$$

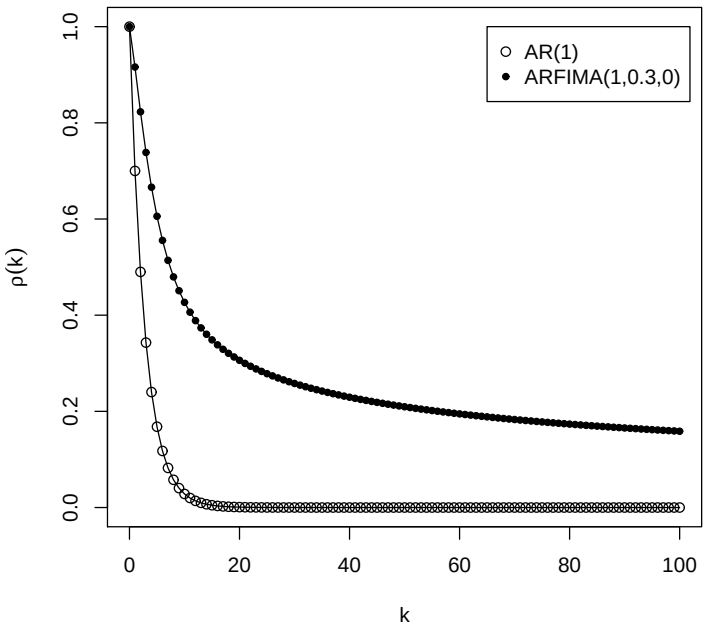
Nakoniec rozdelíme sumu na dve časti pre $|k| > \sqrt{k}$ a $|k| < \sqrt{k}$. Následne použijeme ohraničenie pre γ_1 z kapitoly 1 v zmysle $|\rho(h)| \leq Cr^{-h}$ pre nejaké $C > 0$, $0 < r < 1$ a $k = 0,1,2,\ldots$ Pre odhad γ_0 zase využijeme upravenú rovnicu (2.20). Presný postup, viď [4].

Pre autokorelačnú funkciu potom platí

$$\rho(h) \sim c_\rho |h|^{2d-1}, \quad |h| \to \infty,$$

kde

$$c_\rho = \frac{c_\gamma}{\int_{-\pi}^{\pi} f(\lambda)d\lambda}.$$



Obr. 2.4: Autokorelačná funkcia procesov ARFIMA(1;0,3;0) a AR(1)

Na obrázku 2.4 sú vykreslené autokorelačné funkcie procesov z obrázku 2.1. Priebeh autokorelačných funkcií odpovedá našim doterajším tvrdeniam. U procesu s dlhou pamäťou je klesanie k nule pomalšie - polynomiálne, viď (2.1), oproti exponenciálnemu klesania u ARMA modelov.

2.2.3 Prevod na AR(∞) a MA(∞)

Vo vete 2 sme ukázali, že pre $d \in (-0{,}5; 0{,}5)$ a vhodné polynómy $a(z)$ a $b(z)$ je ARFIMA proces $\{X_t, t \in \mathbb{Z}\}$ stacionárny, kauzálny a ivertibilný. Sformulujme teda vetu, ktorá definuje koefienty pre rozvoj AR(∞) a MA(∞).

Veta 3 (Hosking 1981)**.** *Nech X_t je proces riadiaci sa modelom ARFIMA(0,d,0) a nech $d \in (-0{,}5; 0{,}5)$. Potom platí:*

1. *Proces X_t sa dá vyjadriť ako proces typu AR(∞)*

$$\sum_{k=0}^{\infty} \pi_k X_{t-k} = Y_t,$$

kde Y_t je biely šum a pre π_k platí

$$\pi_k = \frac{\Gamma(k-d)}{\Gamma(k+1)\Gamma(-d)}.$$

Pre $k \to \infty$ platí

$$\pi_k \sim \frac{1}{\Gamma(-d)} k^{-d-1}.$$

2. *Proces X_t sa dá vyjadriť ako proces typu MA(∞)*

$$X_t = \sum_{k=0}^{\infty} \psi_k Y_{t-k},$$

kde Y_t je biely šum a pre ψ_k platí

$$\psi_k = \frac{\Gamma(k+d)}{\Gamma(k+1)\Gamma(d)}.$$

Pre $k \to \infty$ platí

$$\psi_k \sim \frac{1}{\Gamma(d)} k^{d-1}.$$

V [14] môžeme nájsť asymptotické vzorce pre koeficienty π_j a ψ_j obecného procesu ARFIMA(m,d,n) z vety 3

$$\pi_j \sim \frac{a(1)}{b(1)} \frac{j^{-d-1}}{\Gamma(-d)}, \quad j \to \infty,$$

$$\psi_j \sim \frac{a(1)}{b(1)} \frac{j^{d-1}}{\Gamma(d)}, \quad j \to \infty. \tag{2.22}$$

2.2.4 Odhad strednej hodnoty a autokorelačnej funkcie

Pri modeloch ARFIMA taktiež platí, že za odhad strednej hodnoty μ volíme výberový priemer

$$\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Pre rozptyl výberového priemeru potom platí

$$\mathrm{var}(\bar{X}_n) = \frac{1}{n}\left[2\sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\gamma(j)\right) + \gamma(0)\right].$$

Použitím vety 7.1.1 z [4] a faktu, že $\gamma(h)$ konverguje k nule (viď (2.21)) dostávame

$$\mathrm{var}(\bar{X}_n) \to 0, \quad n \to \infty$$

a výsledok

$$\begin{aligned} n\,\mathrm{var}(\bar{X}_n) &\to 0, \quad d \in (-0{,}5; 0), \\ &\to \infty, \quad d \in (0; 0{,}5). \end{aligned}$$

Z predošlého vyplýva, že pre proces s dlhou pamäťou riadiaci sa ARFIMA modelom sa rozptyl nespráva proporčne k n, ako sme zvyknutí pri ARMA procesoch. Ak využijeme asymptotický zápis autokovariančnej funkcie (2.21), tak môžeme odvodiť asymptotický tvar $\bar{X}_n$, viď [16], str. 49

$$\begin{aligned} \mathrm{var}(\bar{X}_n) &\sim 2c_\gamma n^{2d-1}\sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\right)\left(\frac{j}{n}\right)^{2d-1}\frac{1}{n} \\ &\sim 2c_\gamma n^{2d-1}\int_0^1 (1-t)t^{2d-1}dt \\ &\sim \frac{c_\gamma}{d(2d+1)}n^{2d-1}. \end{aligned}$$

Teda pre $d \in (0; 0{,}5)$, kedy u ARFIMA procesu pozorujeme dlhú pamäť, sme potvrdili pozorovanie z úvodu tejto kapitoly o tom, že rozptyl výberového priemeru klesá k nule proporcionálne k $n^{-\alpha}$, kde $\alpha \in (0{,}1)$.

Na záver poznamenajme, že výberový priemer pre procesy s dlhou pamäťou nemusí mať vždy asymptoticky normálne rozdelenie, viď [4].

Pre odhad autokovoriančnej funkcie $\gamma(h)$ využívame momentový odhad

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n), \quad h = 0{,}1, \ldots, n-1,$$

kde $\hat{\gamma}(h) = \hat{\gamma}(-h)$. Odhad autokorelačnej funkcie je potom

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

Oba odhady nie sú nestrannými odhadmi teoretických veličín. Navyše podľa vety 9.1 v [17] platí, že ak je proces lineárny, má konečné štvrté momenty a pre koeficienty ψ_k z lineárneho vyjadrenia platí, že ich súčet absolútne konverguje, tak odhad autokorelačnej funkcie má asymptoticky normálne rozdelenie a platí tzv. Bartlettov vzorec. Z asymptotického vzorca (2.22) pre ψ_k je jasné, že podmienku absolútnej sčítateľnosti splňuje ARFIMA proces len pre $-0{,}5 < d < 0$, čiže keď pozorujeme strednú pamäť. Pre ARFIMA proces s dlhou pamäťou je situácia omnoho komplikovanejšia. Časť výsledkov pre gaussovské procesy je uvedená v knihe *Fox and Taqqu (1986).*

2.3 Predpovede

V tomto odstavci predstavíme v skratke koncept predpovedí náhodných procesov. Budeme uvažovať lineárny invertibilný proces $\{X_t, t \in \mathbb{Z}\}$, t.j. proces, pre ktorý existujú vyjadrenia typu AR(∞) a MA(∞). Model ARFIMA tieto predpoklady spĺňa, viď podkapitola 2.2.3. Teda pre $\{X_t, t \in \mathbb{Z}\}$ platí

$$X_t = \sum_{j=0}^{\infty} \psi_j Y_{t-j}$$

a

$$Y_t = X_t - \sum_{j=1}^{\infty} \pi_j X_{t-j}. \tag{2.23}$$

Proces $\{Y_t, t \in \mathbb{Z}\}$ je biely šum WN(0,σ^2).

Poznámka 5. *Na tomto mieste pripomeňme, že vyjadrenie v tvare MA(∞) procesu $\{X_t, t \in \mathbb{Z}\}$ má vo vete 3 trochu inú podobu. Koeficienty π_j z vety 3 sa však dajú jednoducho prieviesť na tvar zo vzorcu* (2.23)*, ktorý nám v tejto podkapitole bude viac vyhovovať.*

Obecne pod predpoveďou náhodnej veličiny X_{t+h} (predpoveď o h krokov dopredu) na základe minulosti do času t, buď konečnej alebo nekonečnej, rozumieme určitý typ aproximácie, ktorý minimalizuje chybu predpovede od skutočnej hodnoty. Za chybu predpovede sa zvykne používať stredná kvadratická chyba.

Najlepšou predpoveďou je podmienená stredná hodnota, kde v úlohe podmienky stojí minulosť, v tomto prípade reprezentovaná ako Hilberov priestor generovaný pozorovaniami do času t. Obecne sa však podmienená stredná hodnota počíta ťažko, preto ju nahrádzame najlepšou lineárnou predpoveďou.

2.3.1 Nekonečná minulosť

Pri znalosti celej nekonečnej minulosti, ktorá je vyjadrená ako Hilbertov priestor generovaný náhodnými veličinami do častu t, je najlepšou jednokrokovou predpoveďou

$$\hat{X}_{t+1} = \mathsf{E}\left[X_{t+1}|\mathcal{F}_t\right] = \sum_{j=1}^{\infty} \pi_j X_{t+1-j} = \sum_{j=1}^{\infty} \psi_j Y_{t+1-j},$$

pričom stredná štvorcová chyba je $\mathsf{E}\left[X_{t+1} - \hat{X}_{t+1}\right]^2 = \sigma^2$.

Najlepšou lineárnou predpoveďou v čase t o h krokov je potom

$$\hat{X}_t(h) = \mathsf{E}\left[X_{t+h}|\mathcal{F}_t\right] = \sum_{j=0}^{\infty} \pi_j(h) X_{t-j} = \sum_{j=0}^{\infty} \psi_{j+h} Y_{t-j} = \sum_{j=h}^{\infty} \psi_j Y_{t+h-j},$$

pričom pre koeficienty $\pi_j(h)$ platí vzťah, viď [16]

$$\pi_j(h) = \sum_{i=0}^{h-1} \psi_i \pi_{j+h-i}.$$

Pre strednú štvorcovú chybu viackrokovej predpovede potom platí

$$\mathsf{E}\left[X_{t+h} - \hat{X}_t(h)\right]^2 = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2.$$

2.3.2 Konečná minulosť

V praxi samozrejme nepoznáme celú minulosť procesu, poznáme len pozorovania do času t. Najlepšia lineárna predpoveď v čase t o h krokov, viď [17], je potom

$$\tilde{X}_{t+h}(t) = \phi_{t1} X_t + \ldots + \phi_{tt} X_1.$$

Kľúčom k predpovedi zostáva určiť konštanty $\phi_{t1}, \ldots, \phi_{tt}$. Využitím ortogonality v Hilbertovom priestore môžeme zostaviť systém rovníc v maticovom tvare vyjadrený ako

$$\Sigma_t \phi_t = \gamma_{th}, \tag{2.24}$$

kde $\phi_t = (\phi_{1t}, \ldots, \phi_{tt})^T$, $\Sigma_t = [\gamma(i-j)]_{i,j=1,\ldots,t}$ je autokovariančná matica a $\gamma_{th} = (\gamma(t+h-1)\ldots,\gamma(h))^T$.

Koeficienty $\phi_{t1}, \ldots, \phi_{tt}$ môžeme teraz spočítať rekurzívnou metódou tzv. Durbinovým-Levinsonovým algoritmom, ktorý je bližšie predstavený v kapitole 3. Stedná štvorcová chyba pre jednokrokovú predpoveď sa dá pomocou spomínaného algoritmu vyjadriť nasledujúcim spôsobom, viď [16]

$$\nu_t = \mathsf{E}\left[X_{t+h} - \tilde{X}_t(h)\right]^2 = \nu_{t-1}(1 - \phi_{tt}^2).$$

Ak je proces $\{X_t, t \in \mathbb{Z}\}$ stacionárny, tak jednokroková predpoveď v čase t založená na konečnej minulosti $\tilde{X}_{t+1}$ konverguje v norme k predpovedi založenej na minulosti nekonečnej $\hat{X}_{t+1}$ pre t idúce do nekonečna. Zároveň platí, že stredná štvorcová chyba ν_t konverguje k σ^2. To znamená, že ak označíme $\delta_t = \hat{X}_{t+1} - \mathsf{E}\,[\tilde{X}_{t+1}]$, tak platí

$$\sqrt{\delta_t} \to 0, \quad \nu_t \to \sigma^2, \quad t \to \infty.$$

Ak využijeme vlastnosti normy, v našom prípade generovanej strednou hodnotou, a vlastnosti odhadu, tak platí $\delta_t = \nu_t - \sigma^2$. O rýchlosti konvergencie pre proces ARFIMA hovorí nasledujúce tvrdenie.

Tvrdenie 2. *Nech* $\{X_t, t \in \mathbb{Z}\}$ *je ARFIMA(m,d,n), kde* $0 < d < \frac{1}{2}$. *Potom platí*

$$\delta_t \sim \frac{d^2}{t}, \quad t \to \infty. \tag{2.25}$$

Aproximovaná predpoveď

Durbinov-Levinsonov algoritmus môže byť pre výbery s veľkými rozsahmi výpočtovo náročný. Ak použijeme aproximáciu z lemmy 9.1 v [16], ktorá hovorí, že pre dostatočne veľké t platí, že $\phi_{tj} \sim \pi_j$, tak aproximovanú predpoveď na základe konečnej minulosti vytvoríme ako

$$\breve{X}_t(h) = \sum_{j=0}^{t} \pi_j(h) X_{t-j}.$$

Stredná kvadratická chyba takejto predpovede je potom

$$\mathsf{E}\,[X_{t+h} - \breve{X}_t(h)] = \sigma^2 + r_t(h),$$

kde

$$r_t(h) = \mathrm{var}[\sum_{j=t+1}^{\infty} \pi_j(h) X_{t-j}].$$

Pre člen $r_t(h)$ v modeli ARFIMA, za podmienky $\sum_{j=0}^{h-1} \psi_j \neq 0$ a $0 < d < \frac{1}{2}$ platí

$$r_t(h) \sim \left(\sum_{j=0}^{h-1} \psi_j\right)^2 \frac{d \tan(\pi d)}{\pi t}, \quad t \to \infty. \tag{2.26}$$

Všimnime si, že pre jednokrokovú aproximovanú predpoveď, ak d sa blíži k nule, tak zo vťahov (2.25) a (2.26) plynie, že δ_t a r_t majú rádovo podobný priebeh. Naopak ak d sa blíži k $\frac{1}{2}$, tak r_t sa blíži do nekonečna a δ_t je ohraničená, viď [16].

2.4 Aproximácia náhodných procesov s dlhou pamäťou

Výhodnou vlastnosťou procesov, ktoré sa riadia modelmi ARFIMA je, že sa pomocou nich dajú aproximovať procesy s dlhou pamäťou v zmysle nasledujúcej vety. Pre účely tejto vety upravíme definíciu 13 zámenou konštanty c_f za spojitú funkciu f_2, kde $\lim_{\lambda \to 0} f_2(\lambda) = c_2$ a $c_2 \in \mathbb{R}$, na tvar

$$\lim_{\lambda \to 0} \frac{f(\lambda)}{|\lambda|^{-\beta} f_2(|\lambda|)} = 1. \tag{2.27}$$

Veta 4. *Nech $\{Z_t, t \in \mathbb{Z}\}$ je stacionárny lineárny proces s kladnou spektrálnou hustotou f_z, ktorá spĺňa* (2.27)*. Potom existuje taký proces ARFIMA so spektrálnou hustotou f, že pre každé $\epsilon > 0$*

$$\left| \frac{f_z(\lambda)}{f(\lambda)} - 1 \right| < \epsilon,$$

pre $\lambda \in [-\pi, \pi]$.

Dôkaz. V dôkaze sa využíva Stone-Weierstrassova veta, ktorej využitie umožňuje spojitosť hustoty f_z na okolí nuly. Celý dôkaz, viď [16], str. 55.

❑

Veta 4 nám umožňuje procesy, u ktorých detekujeme dlhú pamäť, modelovať procesmi ARFIMA.

Kapitola 3

Odhady parametrov

V tejto kapitole sa zameriame na detekciu dlhej pamäte a odhady jednotlivých parametrov. Existuje mnoho metód, ktorými sa k odhadom dá pristúpiť. My si v ďalšom texte priblížime tie základné.

Na začiatku potrebujeme získať určitý prvotný obraz o priebehu radu. Pre tieto účely sú vhodné heuristické a vizuálne metódy. Neposkytujú nám síce štatisticky hmatateľné výsledky, no pomáhajú nám pri detekcii dlhej pamäte, odhade parametru dlhej pamäte H, resp. d, a následnej voľbe modelu. V prvom rade je dôležité vedieť, čo o danom rade chceme zistiť. Ak nám stačí odhadnúť len dlhodobý vývoj radu, musíme vedieť odhadnúť parameter dlhej pamäte d, resp. H, a napr. parameter c_f z definície 13. V tomto prípade, keď nemáme explicitnú špecifikáciu „krátkej pamäte“, t.j. ARMA štruktúry, hovoríme o semiparametrických metódach. Ak chceme odhadnúť celú korelačnú štruktúru radu a spektrálnu hustotu vo všetkých frekvenciách, potrebujeme poznať ARMA štruktúru radu. Tieto metódy nazývame parametrické.

3.1 Semiparametrické a vizuálne metódy

Na začiatok v skratke uvedieme tri heuristické metódy - *korelogram*, graf rozptylu a *R/S štatistiku*. Tieto metódy umožňujú prvý náhľad na dáta z pohľadu detekcie dlhej pamäte. Okrem spomínaných troch metód existujú ešte mnohé ďalšie - napr. variogram, viď napr. [2]. Na záver uvedieme štatisticky korektnejšiu metódu - regresnú metódu.

3.1.1 Korelogram

Korelogram je graf, ktorý zobrazuje výberové autokorelácie v čase. Pod výberovými autokoreláciami rozumieme ich odhad, viď [5], str. 330. Pripomeňme, že

odhad strednej hodnoty definujeme ako výberový priemer

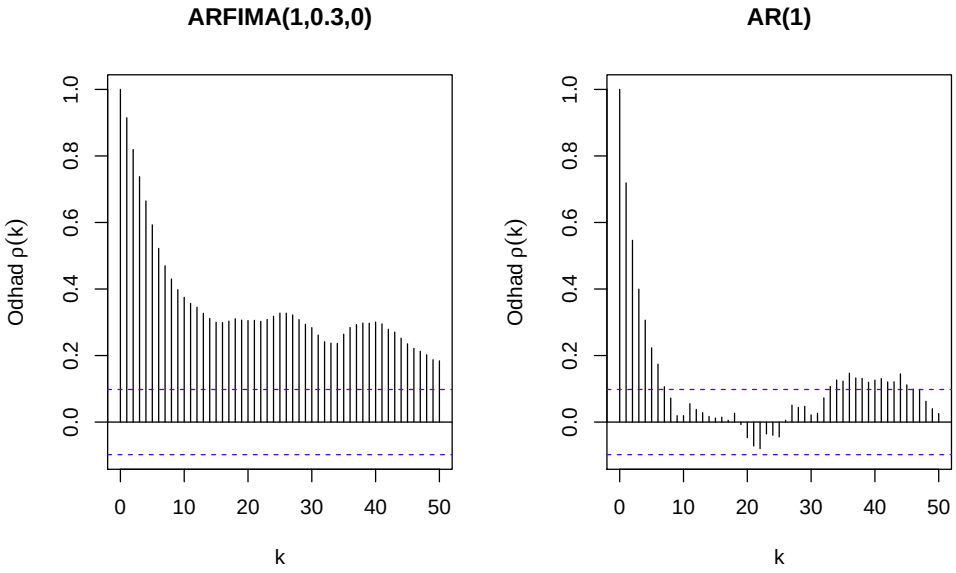
$$\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t.$$

Odhad autokovariančnej funkcie definujeme

$$\hat{\gamma}_k = \frac{1}{n}\sum_{t=k+1}^{n} (X_t - \bar{X})(X_{t-k} - \bar{X}), \quad k = 0, 1, \ldots, n-1$$

a odhad autokorelačnej funkcie ako

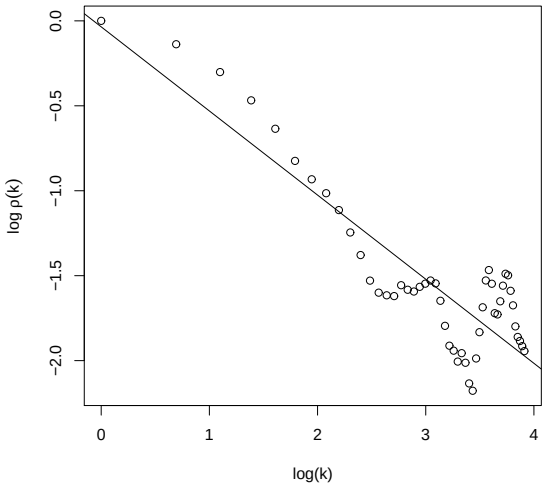
$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}, \quad k = 0, 1, \ldots, n-1.$$



Obr. 3.1: Korelogramy procesu ARFIMA(1;0,3;0) a AR(1)

Korelogram pre procesy z obrázku 2.1 je zobrazený na obrázku 3.1. Ako sme spomínali vyššie pri procesoch s dlhou pamäťou očakávame pomalší - hyperbolický pokles ako u procesov s krátkou pamäťou - exponenciálny pokles.

Dlhá pamäť je asymptotická vlastnosť, preto je potrebné pozerať sa na korelácie pre časovo dosť vzdialené pozorovania, kde ich odhady nemusia byť spoľahlivé. Beran, viď [2], navrhuje používať log-log korelogram vhodný najmä pre rady, u ktorých je vlastnosť dlhej pamäte veľmi významná. Príznakom dlhej pamäte v tomto grafe sú body rozmiestnené okolo priamky so zápornou smernicou (približne $2H - 2$). Pre rady s krátkou pamäťou sú zase body rozmiestnené okolo exponenciálne klesajúcej krivky. Obrázok zobrazuje log-log korelogram procesu ARFIMA z obrázku 2.1. Priamka, ktorá prekladá hodnoty má odhadnutú smernicu $-0{,}49$, teda skutočne blízko k očakávaným $2 * 0{,}8 - 2 = -0{,}4$.



Obr. 3.2: Log-log korelogram procesu ARFIMA(1;0,3;0)

3.1.2 Graf rozptylu

Táto metóda využíva fakt, že pre rozptyl výberového priemeru modelu ARFIMA pre dostatočne veľké n platí

$$\mathrm{var}(\bar{X}_n) \sim cn^{2H-2},$$

pre $c > 0$, viď podkapitola 2.2.4, s využitím vzťahu medzi d a H. Potom, viď [2], môžeme skonštruovať nasledovnú heuristickú metódu.

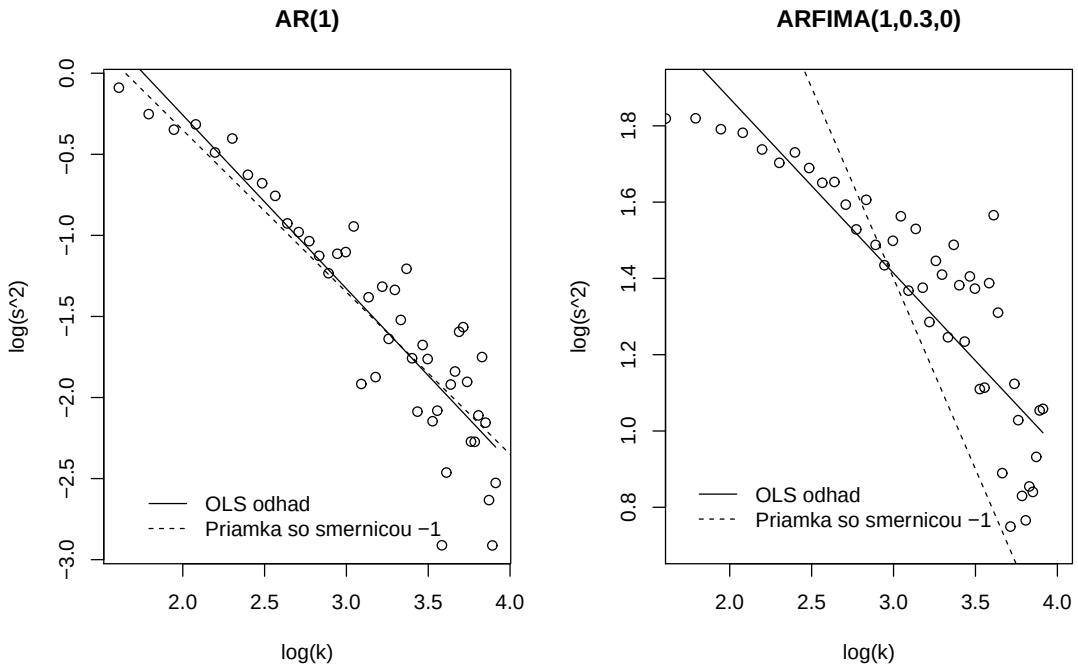
Voľme k prirodzené čísla v rozmedzí $2 \leq k \leq n/2$. Vypočítajme výberové priemery pre podpostupnosti pôvodnej postupnosti o rozsahu k. Nech m_k je počet takto spočítaných rozptylov (malo by byť dostatočne veľké), potom môžeme vypočítané rozptyly označiť $\bar{X}_1(k), \ldots, \bar{X}_{m_k}(k)$. Ak označíme výberový priemer spočítaných priemerov ako $\bar{X}(k)$, kde

$$\bar{X}(k) = m_k^{-1} \sum_{j=1}^{m_k} \bar{X}_j(k),$$

tak môžeme vypočítať výberový rozptyl pre $\bar{X}_1(k), \ldots, \bar{X}_{m_k}(k)$ ako

$$s^2(k) = (m_k - 1)^{-1} \sum_{k=1}^{m_k} (\bar{X}_j(k) - \bar{X}(k))^2.$$

Potom graf rozptylu konštruujeme ako graf $\log s^2(k)$ v závislosti na $\log k$ metódou najmenších štvorcov. Z asymptotického vývoja rozptylu potom platí, že smernica priamky, okolo ktorej sú jednotlivé body grafu rozmiestnené, bude $2H-2$ pre procesy s dlhou pamäťou. Pre procesy s krátkou pamäťou, teda pre $H = 1/2$, resp.



Obr. 3.3: Graf rozptylu procesu AR a ARFIMA

$d = 0$, bude smernica priamky $2 * \frac{1}{2} - 2 = -1$. Graf rozptylu rovnako ako korelogram je užitočný nástroj na prvú diagnostiku modelu, avšak nemá väčšiu štatistickú relevanciu.

Na obrázku 3.3 sme vykreslili graf rozptylu pre procesy AR(1) a ARFIMA(1;0,3;0). Na ľavej časti vidíme proces s krátkou pamäťou, kde smernica priamky zostrojenej pomocou OLS odhadu je $-1{,}07$, teda takmer -1. Na pravej časti je proces s dlhou pamäťou s $d = 0{,}3$, resp. $H = 0{,}8$. Odhad preloženej priamky cez jednotlivé výberové rozptyly má smernicu $-0{,}46$, teda $\hat{H} = 0{,}77$, čiže skoro presne rovné teoretickej hodnote.

3.1.3 *R/S* graf

Táto technika detekcie dlhej pamäte bola objavená hydrológom H. E. Hurstom v roku 1951 v súvislosti s meraním vodného toku Nílu. Rieka Níl na jednej strane spôsobuje dlhodobé záplavy, na strane druhej dlhé suchá. Celkovo jej tok preto pôsobí akoby mal „dlhú pamäť“. Hurst pri snahe regulovať tok Nílu sformuloval nasledujúcu úlohu: Aká má byť kapacita nádrže, aby bola v časoch t až $t+k$ plná, pričom by bol odtok rovnomerný a nádrž by nepretekala? Hurst úlohu obmedzil na diskrétny čas. Ako X_i označme prítok v čase i. Kumulatívny prítok do času j

potom počítame ako

$$Y_j = \sum_{i=1}^{j} X_i.$$

Hurst ukázal, že ideálna kapacita nádrže sa dá vyjadriť ako

$$R(t,k) = \max_{0\le i\le k}\left|Y_{t+1} - Y_i - \frac{i}{k}(Y_{t+k} - Y_i)\right| - \min_{0\le i\le k}\left|Y_{t+1} - Y_i - \frac{i}{k}(Y_{t+k} - Y_i)\right|.$$

Po znormovaní veličinou $S(t,k)$, ktorá je až na konšantu rovná výberovej smerodatnej odchýlke

$$S(t,k) = \sqrt{\frac{\sum_{i=t+1}^{t+k}(X_i - \bar{X}_{t,k})^2}{k}},$$

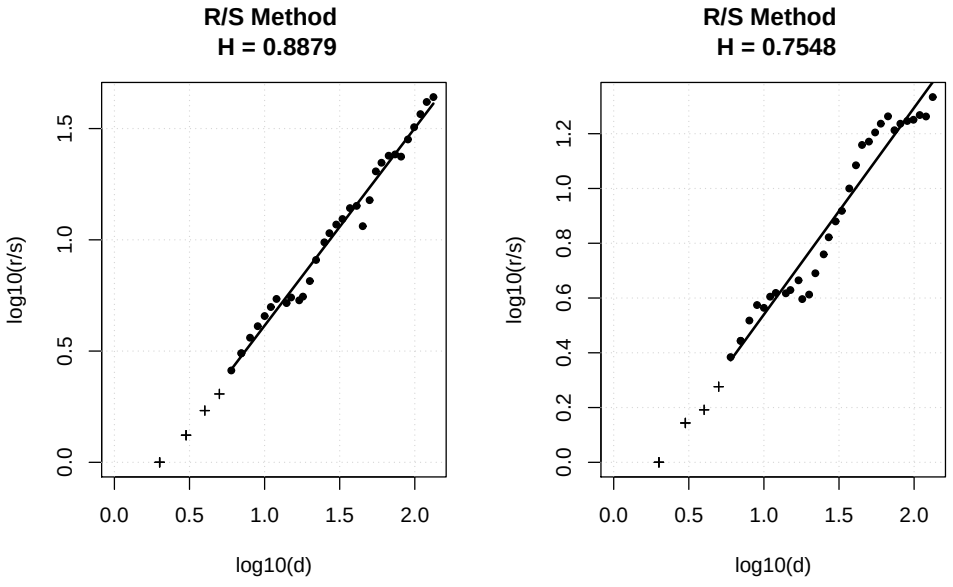
kde

$$\bar{X}_{t,k} = \frac{\sum_{i=t+1}^{t+k} X_i}{k},$$

môžeme definovať známu R/S štatistiku vzťahom,

$$Q(t,k) = R(t,k)/S(t,k). \tag{3.1}$$

Štatistická podstata R/S štatistiky je uvedená napr. v [2]. My na tomto mieste uvedieme len jej aplikáciu. R/S graf potom konštruujeme ako závislosť $\ln[Q(t,k)] \sim \ln[k]$. Pre procesy s krátkou pamäťou pozorujeme, že pre veľké k sú body rozmiestnené okolo priamky so smernicou $\frac{1}{2}$. Naopak táto závislosť pre procesy s dlhou pamäťou vykazuje rozmiestnenie bodov so smernicou $H > \frac{1}{2}$. Teda sme našli odhad parametru dlhej pamäte H.



Obr. 3.4: R/S grafy procesu ARFIMA(1;0,3;0) a AR(1)

Z obrázku 3.4 vidíme, že naozaj sa pri procese s dlhou pamäťou body koncentrujú v okolí priamky so smernicou H v intervale $(0{,}5; 1)$. Parameter H bol odhadnutý metódou najmenších štvorcov na hodnotu 0,70.

R/S štatistika (graf) má rad nedostatkov. Pre veľké k len ťažko spočítame štatistiku Q, taktiež je otázne, či je metóda najmenším štvorcov vhodná. Avšak najväčším problémom R/S štatistiky je jej citlivosť na krátkodobú dynamiku rady. Túto skutočnosť popisuje vo svojom článku Andrew Lo, viď [15]. Ukazuje, že odhad je pre určité procesy s krátkou pamäťou značne vychýlený. Uvádza napríklad autoregresný proces s koeficientom a_1 blízko 0,3. V našom prípade došlo taktiež k značnému znehodnoteniu odhadu pre proces AR(1). Vidíme to hlavne z obrázku 3.4 (pravá časť), kde parameter H bol odhadnutý na 0,75, miesto očakavaného 0,5. Jedným z riešení je podchytiť krátkodobú dynamiku v odhade rozptylu parciálnych súčtov $S(t,k)$, kde okrem rozptylov jednotlivých súm zaradíme aj ich autokovariancie. Zaviedla sa preto tzv. *modifikovaná R/S štatistika*, bližšie k tejto metóde viď [15].

Na druhej strane je R/S graf historicky významný a v praxi často používaný nástroj na prvotnú detekciu, odhad dlhej pamäte.

3.1.4 Regresná metóda

Zlogaritmovaním hustoty (2.3) z definície 13 procesu s dlhou pamäťou a použitím vzťahu (2.4) dostávame

$$\log f(\lambda) \sim \log c_f + (1 - 2H) \log |\lambda|. \tag{3.2}$$

Použitím periodogramu, ktorý je definovaný v kapitole 2, vo Fourierových frekvenciách $\lambda_{j,n} = \frac{2\pi j}{n}$, kde n je rozsah výberu a $j = 1, \ldots, j_0$, pričom j_0 označuje celú časť $(n-1)/2$, a za platnosti (2.17) dostávame

$$\log I(\lambda_{j,n}) \sim \log c_f + (1 - 2H) \log \lambda_{j,n} + \log \eta_j. \tag{3.3}$$

Náhodné veličiny η_j v rovnici (3.3) sú nezávislé exponenciálne náhodné veličiny, pre ktoré platí

$$\mathsf{E}\,(\log \eta_j) = -C,$$

kde $C = 0{,}5572\ldots$ je Eulerova-Mascheroniho konštanta. Presné zdôvodnenie viď [10], vzorec 4.331. Ďalej ak definujeme jednotlivé členy nasledovne (viď [2], str. 96)

$$x_j = \log \lambda_{j,n}, \qquad y_j = \log I(\lambda_{j,n}), \tag{3.4}$$

$$\beta_0 = \log c_f - C, \qquad \beta_1 = 1 - 2H, \tag{3.5}$$

$$\epsilon_j = \log \eta_j + C, \tag{3.6}$$

tak potom môžeme (3.3) prepísať na tvar

$$y_j = \beta_0 + \beta_1 x_j + \epsilon_j. \tag{3.7}$$

Toto je regresná rovnica, kde ϵ_j sú *iid* náhodné veličiny s nulovou strednou hodnotou a rovnakým rozptylom. Parametre β_0,β_1 odhadneme metódou najmenších štvorcov. Odhad parametru H je potom

$$\hat{H} = \frac{1 - \hat{\beta}_1}{2}.$$

Táto semiparametrická metóda má svoje nedostatky. Jedným z nich je predpoklad, že hustota sa správa proporčne k λ^{1-2H} nie len v okolí 0, ale na celom intervale $[-\pi,\pi]$.

Geweke a Porter-Hudak navrhli v [9] metódu, v ktorej pre periodogram používajú len malé frekvencie blízko 0. Výhoda tejto metódy oproti vizuálnym spočíva v tom, že môžeme odvodiť štatisticky uchopiteľné vlastnosti odhadu. Pre model ARFIMA by sa postupovalo analogicky s tým, že vo výpočtoch by sme uvažovali spektrálnu hustotu (2.15).

Predpokladajme teraz, že spektrálna hustota f náhodného procesu má tvar

$$f(\lambda) = f_0(\lambda)[2\sin(\lambda/2)]^{1-2H}, \tag{3.8}$$

čo proces ARFIMA splňuje, viď (2.15). Ak navyše funkcia f_0 je kladná, párna, v 0 má nulovú deriváciu a v okolí 0 má ohraničenú prvú a druhú deriváciu, tak za týchto podmienok môžeme vysloviť vetu, ktorá popisuje asymptotické vlastnosti odhadu $\hat{H}$, viď [2], str. 98.

Veta 5. *Nech $\{X_t, t \in \mathbb{Z}\}$ je stacionárny gaussovský proces so spektrálnou hustotou z* (3.8)*. Nech $\hat{H}$ je odhad parametru H metódou najmenších štvorcov pri frekvenciách λ_j, $l \leq j \leq m$. Ak pre m,l pri $n \to \infty$ platí*

$$m,l \to \infty, \quad \frac{m^5}{n^4} \to 0, \quad \frac{(\log n)^2}{m} \to 0, \quad \frac{l}{m} \to 0, \quad \frac{\sqrt{m}\log m}{l} \to 0,$$

tak

$$\sqrt{m}(\hat{H} - H) \to_d N\left(0, \frac{\pi^2}{24}\right).$$

Dôležitým faktorom je výber hodnôt m a l. Ak je zvolené m príliš malé, tak je odhad vypočítaný z malého počtu hodnôt periodogramu, čo má za výsledok, že odhad $\hat{H}$ je síce menej vychýlený, no má väčší rozptyl. Pri zvolenom väčšom m je odhad kontaminovaný hodnotami periodogramu vysokých frekvencií, čo má za následok vychýlenie odhadu, hoci rozptyl sa zmenší. V praxi sa často odporúča

brať $m = \sqrt{n}$. Rozsah výberu n zohráva takisto významnú rolu. Pri malom rozsahu výberu je odhad $\hat{H}$ regresnou metódou prakticky nevyužiteľný. Beran v [2] uvádza, že pri rozsahu $n = 200$ a voľbe $m = \sqrt{200} \doteq 14$ a $l = 0$, dostávame smerodatnú odchýlku odhadu 0.203 a 95% interval spoľahlivosti $\hat{H} \pm 0.397$, čo je takmer rozsah parametru H, v ktorom pozorujeme dlhú pamäť $\frac{1}{2} < H < 1$.

3.2 Parametrické metódy

Metódy uvedené v tejto podkapitole nám umožnujú odhadovať všetky parametre modelu, t.j. vieme nimi modelovať ako dlhodobý priebeh tak aj krátkodobý. Najznámejšou parametrickou metódou je metóda maximálnej vierohodnosti. My sa v tomto texte obmedzíme na gaussovskú vierohodnostnú funkciu vzhľadom na jej jednoduchosť. Na základe centrálnej limitnej vety (pre procesy generované súčtom nezávislých rovnako rozdelených veličín, viď [17], veta 6.8) sa mnohé z nasledujúcich metód dajú rozšíriť aj na obecné procesy. Existujú mnohé aproximácie gaussovskej vierohodnostnej funkcie. My uvedieme Whittlovu aproximáciu a aproximáciu založenú na autoregresnom rozvoji.

Identifikácia modelu

Pre odhady parametrickými metódami potrebujeme poznať model, t.j. štruktúru krátkodobej dynamiky ARMA. Pri výbere modelu spravidla dochádza k rozhodovaniu, či dáme prednosť presnosti odhadu na úkor príliš veľkého počtu parametrov, alebo vyberieme model s menším počtom parametrov a tým stratíme presnosť.

Medzi najpopulárnejšie metódy identifikácie modelu, ktoré sa snažia vybrať presný model s čo možno najmenším počtom parametrov patria tzv. *informačné kritériá*. Najznámejšie z kritérií je *Akaikeho informačné kritérium*, ktoré obecne definujeme

$$AIC = -2\ln(L) + 2h,$$

kde h určuje počet parametrov v modeli a L je maximalizovaná hodnota vierohodnostnej funkcie odhadnutého modelu. Druhé kritérium má názov *Bayesovo informačné kritérium* a od AIC sa líši len o konštantu

$$BIC = -2\ln(L) + h\ln(n),$$

kde n udáva počet pozorovaní.

Pri procese ARFIMA(m,d,n) sa úloha identifikácie modelu redukuje na optimalizačnú úlohu odhadu parametrov m a n

$$(\hat{m}, \hat{n}) = \arg\min_{(k,l)} A(k,l),$$

kde $A(k,l)$ je vhodné informačné kritérium, viď tiež [5].

Akaikeho informačné kritérium má pri nasledovnej voľbe $A(k,l)$ tvar

$$AIC(k,l) = \ln \hat{\sigma}^2_{k,l} + \frac{2(k+l+2)}{n},$$

pričom $\hat{\sigma}^2_{k,l}$ je odhad rozptylu bieleho šumu z definície modelu ARFIMA a n je dĺžka radu. Všimnime si, že prvý člen penalizuje presnosť, pričom druhý člen penalizuje priveľký počet parametrov. Súčet $k + l + 2$ vyjadruje počet voľných parametrov v modeli, rozptyl a v prípade procesov s dlhou pamäťou aj parameter d. Pre ARMA proces by sme mali $k + l + 1$.

Bayesovo informačné kritérium nesie obdobnú podobou, líšiacu sa len o multiplikatívnu konštantu

$$BIC(k,l) = \ln \hat{\sigma}^2_{k,l} + \frac{(k+l+2)\ln n}{n}.$$

Pri oboch kritériach volíme model s najmenšou hodnotou informačného kritéria. Oba odhady majú svoje pre a proti. Na jednej strane je odhad pri kritériu AIC nekonzistentný, ale eficientný, u kritéria BIC to máme presne naopak, konzistentný odhad s veľkým rozptylom.

3.2.1 Presný odhad metódou maximálnej vierohodnosti

V tejto kapitole budeme uvažovať gaussovský proces $\{X_t, t \in \mathbb{Z}\}$ s dlhou pamäťou, ktorý je stacionárny, kauzálny, lineárny a invertibilný. Predpokladajme, že jeho spektrálna hustota je charakterizovaná m-rozmerným parametrom

$$\theta = (\theta_1, \ldots, \theta_m),$$

t.j. predpokladáme, že spektrálna hustota procesu $\{X_t, t \in \mathbb{Z}\}$ patrí do rodiny hustôt $f(\lambda) = f(\lambda;\theta)$, kde $\theta \in \Theta \subset \mathbb{R}^m$. Ďalej budeme označovať skutočnú hodnotu parametru θ ako θ^0. Pre $X = (X_1, X_2, \ldots, X_n)^T$ definujeme kovariančnú maticu

$$\Sigma_n(\theta) = [\gamma(j-l)]_{j,l=1,\ldots,n}.$$

Determinant matice Σ_n budeme označovať $|\Sigma_n|$.

Združená hustota pre X je potom rovná

$$h(x;\theta) = (2\pi)^{-\frac{n}{2}} |\Sigma_n(\theta)|^{-\frac{1}{2}} e^{-\frac{1}{2}x^T \Sigma_n^{-1}(\theta)x}, \tag{3.9}$$

kde $x = (x_1,x_2,\ldots,x_n)^T \in \mathbb{R}^n$. Funkciu h nazývame vierohodnostná funkcia, v praxi sa však používa jej zlogaritmovaná forma

$$L_n(x;\theta) = \log h(x;\theta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_n(\theta)| - \frac{1}{2}x^T\Sigma_n^{-1}(\theta)x. \qquad (3.10)$$

Odhad parametru θ metódou maximálnej vierohodnosti (MLE - *maximum likelihood estimation*) značíme $\hat{\theta}_n$ a vypočítame ho maximalizovaním logaritmickej vierohodnostnej funkcie $L_n(x;\theta)$. To znamená, že pre $\hat{\theta}_n$ platí

$$L_n'(x;\hat{\theta}_n) = 0,$$

kde $L_n'(x;\theta)$ chápeme ako m-dimenzionálny vektor so zložkami

$$\frac{\partial}{\partial\theta_j}L_n(x;\theta) = -\frac{1}{2}\frac{\partial}{\partial\theta_j}\log|\Sigma_n(\theta)| - \frac{1}{2}x^T[\frac{\partial}{\partial\theta_j}\Sigma_n^{-1}(\theta)]x,$$

pre $j = 1,2,\ldots,m$. Ak označíme maticu druhých derivácií funkcie L_n ako

$$L_n''(x;\theta) = \frac{\partial^2}{\partial\theta_j\theta_l}L_n(x;\theta),$$

pre $j,l = 1,2,\ldots,m$, tak po rozvinutí L_n' do Taylorovej rady dostávame

$$L_n'(x;\hat{\theta}_n) = 0 = L_n'(x;\theta^0) + L_n''(x;\theta^0)(\hat{\theta}_n - \theta^0) + r_n.$$

Ak r_n je asypmtoticky nulový člen rozvoja, tak asymptotické rozdelenie $\hat{\theta}_n - \theta^0$ je rovné asymptotickému rozdeleniu $-[L_n''(x;\theta^0)]^{-1}L_n'(x;\theta^0)$. Dôkaz tohto tvrdenia spolu s ďalsími asymptotickými vlastnosťami MLE odhadu $\hat{\theta}_n$ zhrnutými v nasledujúcej vete sa nachádza v [6].

Veta 6. *Nech $\hat{\theta}_n$ je presný odhad metódou maximálnej vierohodnosti v zmysle* (3.10) *skutočného parametru θ^0, kde $\hat{\theta}_n, \theta^0$ sú m-rozmerné vektory. Potom za podmienok regularity (A0 - A9 v [16], str. 98-99) platí*

- *Konzistencia: $\hat{\theta}_n \to \theta^0$ v pravdepodobnosti, $n \to \infty$.*
- *Eficiencia: $\hat{\theta}_n$ je eficientným odhadom skutočného parametru θ_0 vo Fisherovom zmysle, t.j. Fisherova informačná matica*

  $$\Gamma_n(\theta^0) = \mathsf{E}\,[L_n'(x;\hat{\theta}_n)][L_n'(x;\hat{\theta}_n)]^T$$

  konverguje k inverznej asymptotickej variančnej matici odhadu $\hat{\theta}_n$ (Rao-Cramerova hranica, viď. [1]).

- *Centrálna limitná veta:*

$$\sqrt{n}(\hat{\theta}_n - \theta^0) \to N(0,\Sigma^{-1}(\theta^0)), \quad n \to \infty, \tag{3.11}$$

kde $\Sigma(\theta) = (\Sigma_{ij}(\theta))$, *pričom*

$$\Sigma_{ij}(\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left[\frac{\partial \log f(\lambda;\theta)}{\partial\theta_i}\right]\left[\frac{\partial \log f(\lambda;\theta)}{\partial\theta_j}\right]d\lambda,$$

kde f *je spektrálna hustota procesu.*

Predošlá veta nám teda hovorí, že za určitých podmienok je odhad presnou metódou MLE asymptoticky konzistentný, eficientný a jeho asymptotické rozdelenie je charakterizováné vzťahom (3.11).

Pre model ARFIMA(m,d,n) je vzhľadom na spektrálnu hustotu (2.15) parameter θ tvaru

$$\theta = (\sigma^2,d,a_1,\ldots,a_m,b_1,\ldots,b_n).$$

Presný MLE odhad teda vyžaduje výpočet matice autokovariancií $\Sigma_n(\theta)$, jej determinant $|\Sigma_n(\theta)|$ a inverziu $\Sigma_n^{-1}(\theta)$. Tieto výpočty môžu viesť k časovo a výpočtovo zložitým operáciam. V ďalších podkapitolách si preto ukážeme vybrané techniky na ich výpočet.

Choleského rozklad

Nech Σ je symetrická pozitívne semidefinitná matica, potom platí

$$\Sigma = UU^T,$$

kde U je dolná trojuholníkova matica. Potom pre determinant platí $|\Sigma| = |U|^2 = \prod_{j=1}^{n} u_{jj}^2$, viď napr. [1],[16]. Pre inverziu autokovariančnej matice platí $\Sigma^{-1} = U^{-1}(U^{-1})'$.

Prvky matice U sa počítajú po stĺpcoch zľava a v rámci stĺpca zhora dole. Nižšie sú uvedené explicitné vzorce pre výpočet matice U, pre prvky na diagonále a pre prvky pod ňou:

$$\Sigma_{kk} = \sum_{i=1}^{k} u_{ki}^2 \longrightarrow u_{kk} = \sqrt{\Sigma_{kk} - \sum_{i=1}^{k-1} u_{ki}^2},$$

$$\Sigma_{rc} = \sum_{i=1}^{c} u_{ri}u_{ci} \longrightarrow u_{rc} = \left(\Sigma_{rc} - \sum_{i=1}^{c-1} u_{ri}u_{ci}\right)/u_{cc},$$

kde Σ_{ij} a u_{ij} sú prvky matice Σ, resp. U.

Choleského rozklad je neefektívny pre dlhé časové rady, jeho výpočtová náročnosť je rádovo n^3.

Durbinov-Levinsonov algoritmus

Tento algoritmus zaviedli Durbin a Levinson v roku 1960, bližsie uvedený je v [16]. Základom tohto algoritmu sú jednokrokové $\hat{X}_t$ predpovede založené na konečnej minulosti $\{X_1, \ldots, X_n\}$, kde predpokladáme

$$\hat{X}_1 = 0, \quad \hat{X}_{t+1} = \phi_{t1}X_t + \ldots + \phi_{tt}X_1, \quad pre \quad t = 1, \ldots, n-1.$$

Pričom koeficienty ϕ_{ij} sú definované vzťahmi

$$\begin{aligned}
\phi_{tt} &= [\nu_{t-1}]^{-1}\left[\gamma(t) - \sum_{i=1}^{t-1}\phi_{t-1,i}\gamma(t-i)\right], \\
\phi_{tj} &= \phi_{t-1,j} - \phi_{tt}\phi_{t-1,t-j}, \quad j = 1, \ldots, t-1, \\
\nu_0 &= \gamma(0), \\
\nu_t &= \nu_{t-1}[1 - \phi_{tt}^2], \quad j = 1, \ldots, t-1.
\end{aligned}$$

Ďalej zaveďme chybu predpovede ako $e_t = X_t - \hat{X}_t$, teda $e = (e_1, \ldots, e_n)^T$. Potom, ak L je dolná trojuholníková matica daná vzťahom

$$L = \begin{pmatrix}
1 & & & & & \\
-\phi_{11} & 1 & & 0 & & \\
-\phi_{22} & -\phi_{21} & 1 & & & \\
-\phi_{33} & -\phi_{32} & -\phi_{31} & 1 & & \\
\vdots & \vdots & & & & \\
-\phi_{n-1,n-1} & -\phi_{n-1,n-2} & -\phi_{n-1,n-3} & \cdots & -\phi_{n-1,1} & 1
\end{pmatrix},$$

tak platí $e = LX$. Teraz vieme variančnú maticu Σ rozložiť ako $\Sigma = LDL'$, kde $D = diag(\nu_0, \ldots, \nu_{n-1})$.

Pre determinant autokovariančnej matice platí

$$|\Sigma| = \prod_{j=1}^{n}\nu_{j-1}.$$

Pre člen obsahujúci inverznú maticu zo vzťahu (3.10) platí

$$x'\Sigma^{-1}x = e'D^{-1}e.$$

Potom môžeme logaritmickú vierohodnostnú funkciu (3.10) zapísať v tvare

$$L(\theta) = -\frac{1}{2}\sum_{t=1}^{n}\log\nu_{t-1} - \frac{1}{2}\sum_{t=1}^{n}\frac{e_t^2}{\nu_{t-1}}.$$

Výpočtová zložitosť tohto algoritmu pre procesy ARFIMA je n^2.

Výpočet autokovariančnej funkcie

Pre využitie Choleského rozkladu a Durbinovho-Levinsonovho algoritmu potrebujeme vedieť efektívne rýchlo počítať autokovariancie. S veľkou dávkou presnosti sa dá odhadovať autokovariančná funkcia nasledujúcim spôsobom, viď [16].

Táto metóda spočíva v rozklade ARFIMA procesu na ARMA časť a na jeho frakcionálne integrovanú časť FI. Označme $\gamma_1(\cdot)$ autokovariančnú funckiu ARMA časti a autokovariančnú funkciu frakcionálneho šumu FI $\gamma_2(\cdot)$, pričom využívame tvrdenie 1. Obe autokovariančné funkcie s explicitným vzorcom sa nachádzajú v kapitole 2. Potom definujme novú funkciu

$$\gamma(h) = \sum_{j=-\infty}^{\infty} \gamma_1(j)\gamma_2(j-h). \tag{3.12}$$

Spôsob, akým je definovaná funkcia $\gamma(\cdot)$, nazývame *konvolúcia* funkcií γ_1, γ_2. Ak pre nejaké m usekneme nekonečnú sumu v (3.12), tak dostávame aproximáciu autokovariančnej funckie

$$\gamma(h) \sim \sum_{j=-m}^{m} \gamma_1(j)\gamma_2(j-h).$$

3.2.2 Whittlova aproximácia MLE

Asi najznámejšou aproximáciou gaussovskej logaritmickej funkcie je Whittlova aproximácia, ktorá ponúka veľmi rýchly odhad parametrov. Uvažujme znovu gaussovský vektor $X = (X_1, \ldots, X_n)$, ktorý má nulovú strednú hodnotu a kovariančnú maticu $\Sigma(\theta)$. Zlogaritmovanú vierohodnostnú funkcu (3.10) vydelíme rozsahom výberu n pre zjednodušenie ďalších zápisov

$$\frac{1}{n}L_n(x;\theta) = \frac{1}{n}\log h(x;\theta) = -\frac{1}{2}\log 2\pi - \frac{1}{2n}\log|\Sigma(\theta)| - \frac{1}{2n}x^T\Sigma^{-1}(\theta)x. \tag{3.13}$$

V (3.13) sú dva členy závislé na parametri θ, ktoré sa budeme snažit aproximovať

$$\frac{1}{2n}\log|\Sigma(\theta)| \quad a \quad \frac{1}{2n}x^T\Sigma^{-1}(\theta)x. \tag{3.14}$$

Prvý člen (3.14) sa dá aproximovať (viď [12]) ako

$$\frac{1}{n}\log|\Sigma(\theta)| \to \frac{1}{2\pi}\int_{-\pi}^{\pi}\log[2\pi f(\lambda;\theta)]d\lambda, \quad pre \quad n \to \infty. \tag{3.15}$$

Druhý člen (3.14) môžeme aproximovať nasledujúcim spôsobom, viď [16]

$$\frac{1}{2n}x^T\Sigma^{-1}(\theta)x \sim \sum_{l=1}^{n}\sum_{j=1}^{n} x_l \left\{ \frac{1}{8\pi^2 n}\int_{-\pi}^{\pi} f^{-1}(\lambda;\theta)\exp[i\lambda(l-j)]d\lambda \right\} x_j$$
$$= \frac{1}{8\pi^2 n}\int_{-\pi}^{\pi} f^{-1}(\lambda;\theta)\sum_{l=1}^{n}\sum_{j=1}^{n} x_l x_j \exp[i\lambda(l-j)]d\lambda$$
$$= \frac{1}{8\pi^2 n}\int_{-\pi}^{\pi} f^{-1}(\lambda;\theta)\left|\sum_{j=1}^{n} x_j \exp(i\lambda j)\right|^2 d\lambda$$
$$= \frac{1}{4\pi}\int_{-\pi}^{\pi}\frac{I(\lambda)}{f(\lambda;\theta)}d\lambda, \tag{3.16}$$

kde $I(\lambda)$ je periodogram definovaný v kapitole 2. V tomto okamihu môžeme vierohodnostnú funkciu (3.13) aproximovať pomocou (3.15) a (3.16) a zaviesť novú vierohodnostnú funkciu L_1 vzťahom

$$L_1(\theta) = -\frac{1}{4\pi}\left[\int_{-\pi}^{\pi}\log f(\lambda;\theta)d\lambda + \int_{-\pi}^{\pi}\frac{I(\lambda)}{f(\lambda;\theta)}d\lambda\right]. \tag{3.17}$$

Na výpočet L_1 je potrebý výpočet integrálov, ktorý môže byť pre zjednodušenie nahradený Riemannovým súčtom, viď [16] a [2].

$$\int_{-\pi}^{\pi}\log f(\lambda;\theta)d\lambda \sim \frac{2\pi}{n}\sum_{j=1}^{n}\log f(\lambda_j;\theta), \tag{3.18}$$

$$\int_{-\pi}^{\pi}\frac{I(\lambda)}{f(\lambda;\theta)}d\lambda \sim \frac{2\pi}{n}\sum_{j=1}^{n}\frac{I(\lambda_j)}{f(\lambda_j;\theta)}, \tag{3.19}$$

kde $\lambda_j = \frac{2\pi j}{n}$ sa nazývajú Fourierove frekvencie. Na záver môžeme uviesť diskrétnu podobu zlogaritmovanej vierohodnostnej funkcie L_1 z (3.17)

$$\tilde{L}_1(\theta) = -\frac{1}{2n}\left[\sum_{j=1}^{n}\log f(\lambda_j;\theta) + \sum_{j=1}^{n}\frac{I(\lambda_j)}{f(\lambda_j;\theta)}\right].$$

Dahlhaus v [6] ukazuje, že odhad Whittlovou metódou $\hat{\theta}_n$ je za určitých podmienok regularity konzistentný a $\sqrt{n}(\hat{\theta}_n - \theta^0) \to N(0,\Sigma^{-1}(\theta^0))$, pre $n \to \infty$.

Výhodou Whittlovej aproximácie je, že sa dá použiť aj pri negaussovských procesoch. Ak znormalizujeme spektrálnu hustotu $f(\lambda;\theta)$ tak, aby

$$\int \log f(\lambda;\theta)d\lambda = 0,$$

tak môžeme (3.17) prepísať na tvar

$$\bar{L}_1(\theta) = -\frac{1}{4\pi}\int_{-\pi}^{\pi}\frac{I(\lambda)}{f(\lambda;\theta)}d\lambda \tag{3.20}$$

s analogickou diskrétnou verziou. Dá sa ukázať, že ak $\{X_t, t \in \mathbb{Z}\}$ je lineárny a stacionárny proces, ktorý sa dá zapísať v tvare

$$X_t = \sum_{j=0}^{\infty} \psi_j(\theta) Y_{t-j},$$

kde Y_t je *iid*, pričom platí $\sum_{j=0}^{\infty} \psi_j^2(\theta) < \infty$, tak odhad Whittlovou aproximáciou $\hat{\theta}_n$, ktorý maximalizuje $\bar{L}_1$ v (3.20) je konzistentný a spĺňa podmienku normality v zmysle limitného rozdelenia $\sqrt{n}(\hat{\theta}_n - \theta^0)$, viď [16], str. 81. Ako sme ukázali v dôkaze vety 2, pre ARFIMA procesy je podmienka $\sum_{j=0}^{\infty} \psi_j^2(\theta) < \infty$ splnená.

3.2.3 Autoregresná aproximácia MLE

Výpočet presného MLE býva náročný, dá sa preto využiť aproximácia využívajúca autoregresný rozvoj AR(∞). Nech $\{X_t, t \in \mathbb{Z}\}$ je proces s dlhou pamäťou s autoregresným rozvojom

$$X_t = Y_t + \pi_1(\theta) X_{t-1} + \pi_2(\theta) X_{t-2} + \pi_3(\theta) X_{t-3} + \ldots, \tag{3.21}$$

kde $\{Y_t, t \in \mathbb{Z}\}$ je WN(0,σ^2). Pre model ARFIMA sú parametre π_j uvedené v kapitole 2, ktoré vzhľadom k poznámke 5 uvádzame v tvare (2.23). V praxi však máme k dispozícii len konečný počet pozorovaní $X_1, X_2, \ldots, X_n$, preto uvažujeme model

$$X_t = \tilde{Y}_t + \pi_1(\theta) X_{t-1} + \pi_2(\theta) X_{t-2} + \ldots + \pi_m(\theta) X_{t-m}, \tag{3.22}$$

pre $m < t \leq n$. Potom definujme vierohodnostnú funckiu L_2 ako

$$L_2(\theta) = \sum_{t=m+1}^{n} [X_t - \pi_1(\theta) X_{t-1} - \pi_2(\theta) X_{t-2} - \ldots - \pi_m(\theta) X_{t-m}]^2. \tag{3.23}$$

Minimalizovaním tejto funkcie dostávame aproximatívny odhad $\hat{\theta}_n$. Všimnime si, že k odhadu parametru θ využívame najlepšiu lineárnu predpoveď založenú len na minulých hodnotách. Táto metóda sa niekedy označuje aj metóda podmienených najmenších štvorcov, viď [17].

Existuje viacero vylepšení autoregresnej aproximácie MLE, napr. Beranova metóda, Haslettova-Rafteryho metóda, atď. My si uvedieme Beranovu metódu, ktorá je popísaná v [2].

Ak predpokladáme, že rad $\{X_t, t \in \mathbb{Z}\}$ je gaussovský, tak môžeme pre vektor $Y = (Y_1, \ldots, Y_n)$ z (3.21) definovať logaritmickú vierohodnostnú funkciu $\tilde{L}_3$ vzťahom

$$\tilde{L}_3(\theta) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2} \sum_{t=1}^{n} \left(\frac{y_t}{\sigma} \right)^2.$$

Keďže v praxi hodnoty pre pre $t < 0$ nepozorujeme, môžeme definovať rad U_t ako odhad $Y_2, \ldots, Y_n$

$$U_t = X_t - \sum_{j=1}^{t-1} \pi_j(\theta) X_{t-j},$$

pre $t = 2, \ldots, n$. Definujme ďalej $R_t(\theta) = U_t(\theta)/\sigma$. Pre model ARFIMA($m$,$d$,$n$) je $\theta =$ (σ^2,d,$a_1, \ldots$,a_m,$b_1, \ldots$,b_n). Po dosadení do $\tilde{L}_3$ dostávame maximálne vierohodný odhad $\hat{\theta}_n$ minimalizovaním

$$2n \log \sigma + \sum_{t=2}^{n} R_t^2(\theta).$$

Ak predošlý vzťah parciálne zderivujeme podľa θ, môžeme minimalizáciu previesť na problém riešenia systému nelineárnych rovníc

$$\sum_{t=2}^{n} \{R_t(\theta)\dot{R}_t(\theta) - E[R_t(\theta)\dot{R}_t(\theta)]\} = 0, \tag{3.24}$$

kde $\dot{R}_t(\theta) = \left(\frac{\partial R_t(\theta)}{\partial \theta_1}, \ldots, \frac{\partial R_t(\theta)}{\partial \theta_r}\right)^T$. Beran v [2] ukazuje, že odhad $\hat{\theta}_n$ vypočítaný ako riešenie sústavy (3.24) je konzistentný, eficientný a $\sqrt{n}(\hat{\theta}_n - \theta^0) \to N(0,\Sigma^{-1}(\theta^0))$, pre $n \to \infty$.

Kapitola 4

Analýza dát

V tejto kapitole predstavíme vybrané balíčky štatistického softvéru R, ktoré riešia problematiku časových radov s dlhou pamäťou. Následne si ukážeme reálny príklad dát s dlhou pamäťou, na ktorých predvedieme aplikáciu jednotlivých balíčkov. Na spracovanie dát budeme požívať štatistický softvér R, viď [18].

4.1 Vybrané balíčky R

V tejto podkapitole uvedieme základné balíčky z programu R, ktoré sa zaoberajú procesmi s dlhou pamäťou. Ukážeme ich základné funkcie a vlastnosti. Konkrétne zadávanie funkcií v R sa dá nájsť na priloženom disku v rámci analýzy dát z Nílu.

4.1.1 Balíček *longmemo*

Kompletná dokumentácia tohto balíčku sa nachádza v [3]. Balíček pozostáva zo zozbieraných algoritmov, ktoré zaviedol a popísal vo svojich prácach Jan Beran. Časť algoritmov priamo s kódom z R je uvedený v [2]. Tento balík taktiež obsahuje zbierku dát, ktoré slúžia ako modelové príklady procesov s dlhou pamäťou. Dajú sa tu nájsť popri dátach *VideoVBR* (počet kódovaných informácií na snímkoch video sekvencií), *NhemiTemp* (mesačné teploty severnej pologule v rokoch 1854 - 1989), a.i., aj dáta *NileMin*, ktoré obsahujú minimálne hladiny Nílu v rokoch 622 až 1281. Práve týmito dátami sa budeme zaoberať pri aplikovaní rôznych balíčkov, viac v podkapitole 4.2.

WhittleEst

WhittleEst() je užitočná funkcia, ktorá vypočíta odhad parametrov metódou maximálnej vierohodnosti za využitia Whittlovej aproximácie, viď podkapitola 3.2.2. Na tomto mieste a ani v ďalšom texte nebudeme uvádzať úplný výčet argumentov danej funkcie, uvedieme len tie, ktoré sú pre túto prácu zaujímavé.

Argumentom `model` si vyberieme medzi ARFIMA modelom a gaussovským frakcionálnym šumom. Pre určenie rádov m, resp. n modelu ARFIMA(m,d,n) slúžia argumenty `p` a `q`. Po zavolaní funkcie dostaneme kompletný odhad autoregresných parametrov, parametrov kĺzavých súčtov a parametru dlhej pamäte $H = d + 0{,}5$ s odhadnutými smerodatnými odchýlkami. Aplikáciou funkcie *plot()* v spojitosti s *lines()* na odhad funkciou *WhittleEst()* dostaneme užitočný graf, na kotorom je periodogram a odhadnuté spektrum Whittlovou aproximáciou. Na druhej strane nám u tejto funkcie chýbajú vypočítané reziduá a fitted values.

Ostatné

Balíček *longmemo* obsahuje ďalšie zaujímavé funkcie. Pod funkciou *ckARMA0()* je priamo podľa vzorca (2.18) implementovaný výpočet autokovariančnej funkcie procesu ARFIMA(0,d,0). Na výpočet spektrálnej hustoty procesu ARFIMA(m,d,n) slúži *specARIMA()*. Viac o uvedených a ďalších funkciách viď [3].

4.1.2 Balíček *arfima*

Balíček *arfima* patrí medzi základné pre analýzu dlhej pamäte, ktorého dokumentáciu môžeme nájsť v [19].

arfima.sim

Začnime simuláciou radu. Funkcia *arfima.sim()* generuje ARIMA proces, v ktorom sa dá voliť medzi rôznymi typmi dlhej pamäte (ARFIMA, frakcionálny gaussovský šum,...). Pre model ARFIMA(m,d,n) musíme definovať argument `dfrac` nami požadovanou hodnotou d. Argumenty `phi`, resp. `theta`, definujú parametre $a_1, \ldots, a_m$, resp. $b_1, \ldots, b_n$. Pre nastavanie celočíselnej diferencie, napr. na stacionarizáciu radu, slúži paramater `dint`. Balíček *arfima* pracuje vo všetkých funkciách s modelom ARFIMA(m,d,n) v tvare

$$\phi(B)(1-B)^d X_t = \theta(B) Y_t,$$

kde Y_t je biely šum a operátory $\phi(B)$ a $\theta(B)$ sú na rozdiel od operátorov $a(B)$ a $b(B)$ z definície 15 definované s opačnými znamienkami. To znamená, že $\phi(B) =$

$1 - \phi_1 B - \ldots - \phi_m B^m$, a to isté platí aj pre operátor $\theta(B)$. Na túto skutočnosť je potrebné dávať obzvlášť pozor pri interpretácii odhadov, ktoré uvedieme neskôr. Simulovaný rad sa nakoniec generuje ako výber z normálneho rozdelenia, ktorý má autokovariančnú štruktúru definovanú zadanými parametrami.

arfima

Hlavnou funkciou tohto balíčku je *arfima()*, ktorá počíta presný odhad metódou maximálnej vierohodnosti. Na vstupe sa dá voliť z množstva parametrov. Hlavným je typ dlhej pamäte, ktorým chceme náš rad simulovať, pre model ARFIMA je to argument `lmodel` s hodnotou `"d"`. Argument `order` určuje ARIMA štruktúru radu a zadáva sa v tvare `c(p,d,q)`, t.j. `p,q` udávajú rády štruktúry ARMA a `d` udáva celočíselný rád diferencovania. Argumentom `fixed` môžeme dopredu určiť konkrétne hodnoty parametrov $a_1, \ldots, a_m, b_1, \ldots, b_n$. Funkcia *arfima()* dokáže podchytiť aj sezónnosť argumentom `seasonal`. Taktiež je v ponuke vybrať si v argumente `whichopt` optimalizačný algoritmus na riešenie nelineárných rovníc pri výpočte MLE. Autormi balíčku je odporúčaný a prednastavený algoritmus BFGS (Broyden-Fletcher-Goldfarb-Shanno), viac o ich kladoch a záporoch viď [19].

Výstupom je odhad modelu dátového typu *arfima*, pričom je vypočítaných vždy viacero riešení, tzv. *modes*, ktoré sú zoradené podľa vypočítanej logaritmickej vierohodnostnej funkcie `logl`. K odhadom sú vypočítané smerodatné odchýlky. Riešenia, ktoré sa blížia k hranici nesplnenia podmienok invertibility alebo stacionarity, sú vyznačené hviezdičkami.

K odhadnutým hodnotám radu (tzv. fitted values), resp. reziduám pristupujeme funkciou *fitted()*, resp. *resid()*, na vybrané riešenie, pričom fitted values sú očistené o odhad strednej hodnoty. K samotnému vykresleniu fitted values do grafu pôvodného radu je teda nutné pripočítať strednú hodnotu.

Ostatné

Takisto ako v balíčku *longmemo* aj tento balíček obsahuje výpočet teoretickej autokovariančnej funkcie pre zadaný model *tacvfARFIMA()*, funkcia *tacvf()* vypočíta autokovariančnú funkciu pre odhadnutý model typu *arfima*. Nájdeme tu aj implementované informačné kritéria *AIC()* a *BIC()* na model typu *arfima*. Predpovede sa dajú skonštruovať pomocou príkazu *predict()*, argument `n.ahead` určí počet krokov predpovede.

4.1.3 Balíček *fracdiff*

Tento balíček je vyložene vhodný pre prácu s frakcionálne diferencovanými radmi. Podrobne o jeho funkčnosti viď [8].

fracdiff

Funkcia *fracdiff()* konštruuje odhad metódou maximálnej vierohodnosti s tým, že pre vierohodnostnú funkciu používa Haslettovu-Rafteryho aproximáciou vierohodnostnej funkcie, viď [16], str. 72. Haslettova-Rafteryho aproximácia sa radí medzi autoregresné aproximácie vierohodnostnej funkcie. Vstupnými argumentmi sú `nar`, resp. `nma`, ktoré predstavujú rády m, resp. n modelu ARFIMA(m,d,n). Argumentom `drange` si môže užívateľ nastaviť interval pre d, vzhľadom ku ktorému bude maximalizovaná vierohodnostná funkcia. Výsledkom je odhad všetkých parametrov so smerodatnými odchýlkami. Aj v tomto prípade si treba dať pozor na správne odčítanie odhadnutých parametrov, pretože aj tu sa využíva konvecnia definovania modelu ARFIMA s obrátenými znamienkami.

fdGPH

Tento príkaz počíta odhad parametru dlhej pamäte d autoregresnou metódou, viď kapitola 3.2.3. Výsledkom je odhad d s asymptotickou a štandardnou smerodatnou odchýlkou odhadu d. Voliteľným argumentom pri zadávaní príkazu je `bandw.exp`, ktorým môžeme určiť parameter m z vety 5. Argumentom je exponent v zmysle $m = n^{bandw.exp}$, kde n je rozsah dát a za m sa vezme dolná celá časť.

Ostatné

Funkcia *diffseries()* je užitočný nástroj na diferencovanie rady aj neceločíselného rádu, viď vzorec (2.9), pre rozvoj pomocou gamma funkcie. Takisto aj tu nájdeme príkaz *fracdiff.sim()* pre simuláciu ARFIMA procesu.

4.1.4 Balíček *forecast*

Tento obsiahly balíček zachycuje celú paletu analýzy časových radov. Nezameriava sa len na procesy s dlhou pamäťou, napriek tomu v ňom vieme nájsť využiteľné funkcie. Dokumentácia tohto balíčku viď [13].

arfima

Funkcia *arfima()* slúži pre odhad parametrov v modeli ARFIMA. Na rozdiel od ostatných balíčkov disponuje argumentom `estim`, ktorým volíme medzi presným MLE a Haslettovou-Rafteryho aproximáciou. Táto procedúra je spojením procedúr *fracdiff()* z rovnomenného balíčku a procedúry *auto.arima()* priamo z balíčku *forecast*. Na rozdiel od predchádzajúcich procedúr je tento balíček výhodný v tom, že pre dáta automaticky zvolí parametre m a n. Algoritmus najprv vypočíta odhad parametru d z modelu ARFIMA(2,d,0), potom je pôvodná rada zdiferencovaná odhadnutou hodnotou d, následne sa na zdiferencovaný rad aplikuje procedúra *auto.arima()*, ktorá na základe informačných kritérií zvolí vhodné m a n. Na záver sa znovu odhadne model.

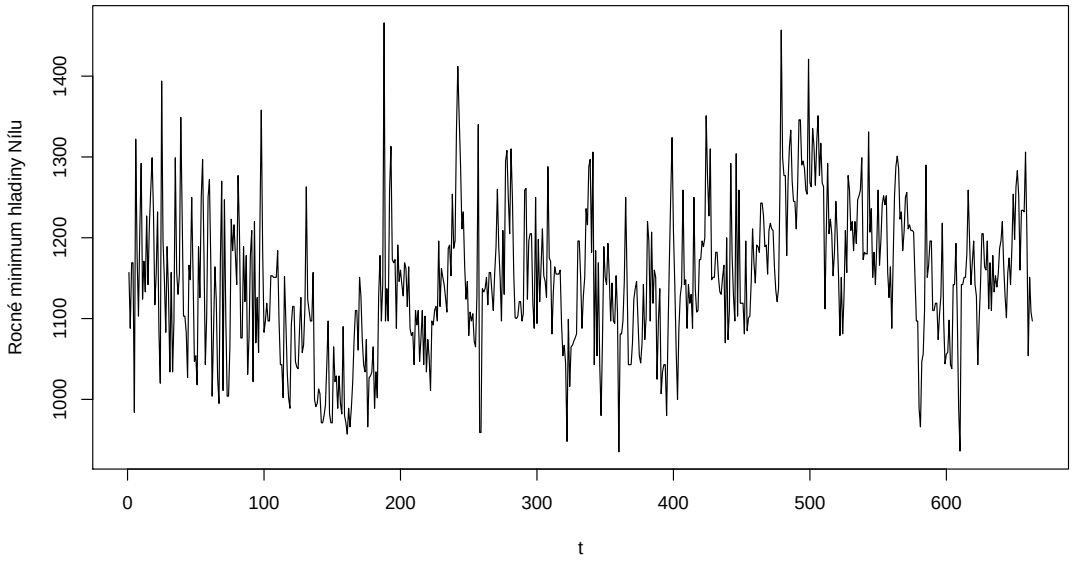
Predpoveď v tomto balíčku sa volá príkazom *forecast()* a poskytuje ako intervalovú tak aj bodovú predpoveď.

4.1.5 Balíček *fArma*

Tento balíček využíva už popísané funkcie v iných balíčkoch. Jeho hlavnou výhodou je, že zjednocuje viacero balíčkov dohromady. Ďalším pozítivom je kvalitná a podrobná dokumentácia, viď [7]. Funkcia, ktorá nás zaujala, *rsFit()*, počíta odhad Hurstovho parametru R/S metódou, viď podkapitola 3.1.3. Obsahuje užitočný voliteľný paramater `doplot`, ktorý umožní grafické znázornenie R/S štatistiky, viď obrázky 3.4, 4.3.

4.2 Dáta z rieky Níl

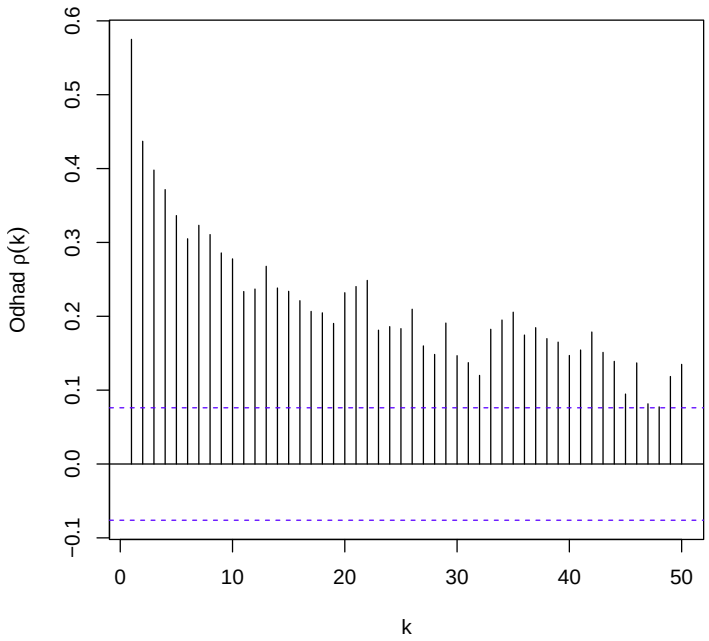
Skúmanie hladiny rieky Níl bolo už sčasti spomenuté pri zavádzaní R/S grafu v podkapitole 3.1.3. Zdrojom dát je balíček *longmemo*, ktorý obsahuje 663 pozorovaní z rokov 622 až 1281, viď obrázok 4.1. Hodnoty vyjadrujú minimálnu výšku hladiny Nílu za daný rok. Dáta boli namerané na ostrove Roda neďaleko od Káhiry a uverejnené v publikácii od Toussona v roku 1925. Práve tieto dáta dali prvotný impulz k skúmaniu dlhodobej závislosti medzi jednotlivými pozorovaniami. Týmito dátami sa zaoberal už spomínaný hydrológ H. E. Hurst, v spojitosti s čím vzniklo pomenovanie Hurstov parameter H. Na základe týchto pozorovaní zaviedol B. B. Mandelbrot nový druh procesov, tzv. frakcionálny gaussovský šum.



Obr. 4.1: Priebeh nameraných miním na rieke Níl v jednotlivých rokoch

4.2.1 Grafická analýza

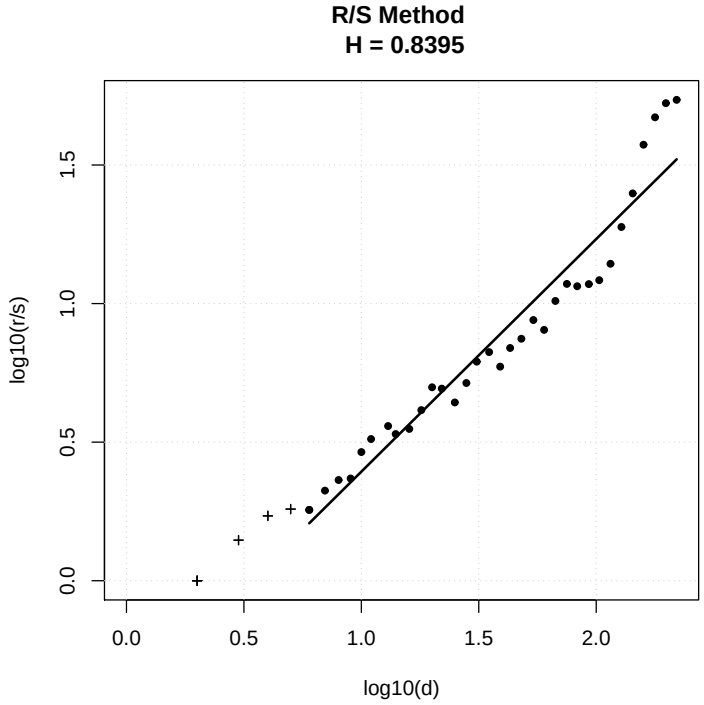
V podkapitole 3.1 sme predstavili grafické metódy, ktorými vieme prvotne detekovať proces s dlhou pamäťou. O tom, že naše dáta majú „dlhú pamäť“ na-



Obr. 4.2: Korelogram dát z rieky Níl

svedčuje ich korelogram zobrazený na obrázku 4.2. Je dobre viditeľné, že korelácie klesajú k nule pomaly. Máme dojem, že polynomiálne.

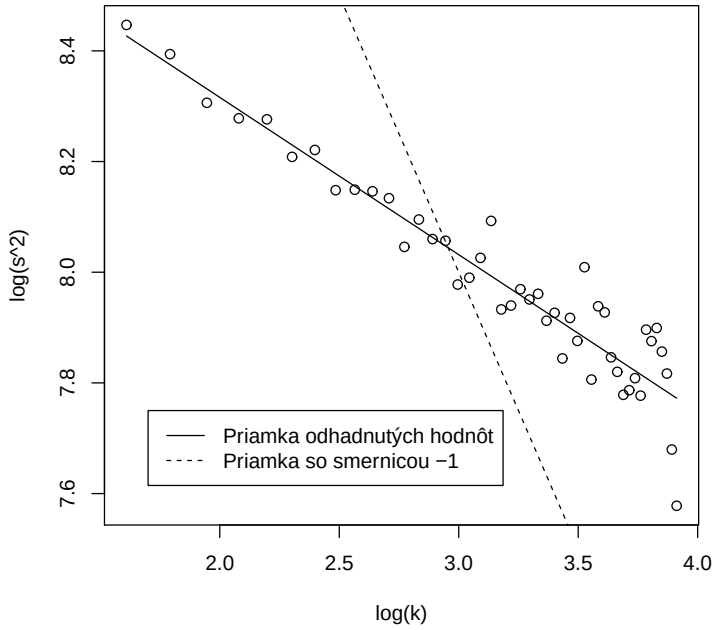
Ďalším vodítkom pre výber modelu, ktorý zohľadňuje dlhododobú závislosť medzi pozorovaniami, je R/S graf. Na jeho vykreslenie a vypočítanie odhadu Hurstovho parametru H sme použili balíček *fArma* a funkciu *rsFit()*. Z obrázku 4.3 vidíme, že smernica priamky je približne 0,84, čo je v tomto prípade rovno hodnota H a tá spadla do intervalu $(0{,}5; 1)$. Teda máme ďalší dôvod pre model s dlhou pamäťou.



Obr. 4.3: R/S graf z dát z rieky Níl

Ďalej sa pozrime na priebeh rozptylu výberového priemeru. V programe R sme skonštruovali algoritmus pre graf rozptylu, ktorý je popísaný v podkapitole 3.1.2. Jednotlivé body sme preložili priamkou pomocou metódy najmenších štvorcov. Intercept vyšiel 8,88 a smernica $-0{,}28$, ktorá vieme, že je rovná $2H - 2$. Teda dospievame k záveru, že dáta vykazujú dlhú pamäť s Hurstovým parametrom 0,86, teda veľmi blízko k odhadu pomocou R/S metódy.

Na záver sa pozrime na regresnú metódu odhadu parametru dlhej pamäte d z podkapitoly 3.1, ktorá síce nepatrí do grafickej analýzy, no patrí do triedy semiparametrických metód. Funkcia *fdGPH()* z balíčku *fracdiff* počíta odhad upravenou regresnou metódou podľa Gewekeho a Porter-Hudaka, ako sme uviedli v podkapitole 3.1.4. V zmienenej podkapitole sme sa rovnako venovali diskusii



Obr. 4.4: Graf rozptylu pre dáta z Nílu

o voľbe parametru m. Pozrime sa na jednotlivé odhady pri rôznych voľbách m. Prednastavenou hodnotou argumentu `bandw.exp` je 0,5, teda $m = 25$. Pri takejto voľbe dostávame odhad $\hat{d} = 0{,}503$ so smerodatnou odchýlkou 0,14, ktorý je evidentne nepoužiteľný. Pri zvýšení m na 94, t.j. `bandw.exp` = 0,7, dostávame odhad $\hat{d} = 0{,}396$, ktorý môže byť v porovnaní s predošlým odhadom vychýlenejší, no s menším rozptylom 0,08. Z vypočítaného vidíme, že rozumenjšou voľbou m je druhá možnosť, v ktorej odhad parametru d je, ako uvidíme neskôr, blízko odhadu metódou presnej maximálnej vierohodnosti.

4.2.2 Odhady parametrov v programe R

Pre modelovanie dát s dlhou pamäťou volíme model ARFIMA(m,d,n), kde autoregresné parametre, parametre kĺzavých súčtov a d odhadneme programom R. Rády m a n si musíme zvoliť sami. Odhadneme preto najprv modely postupne pre $m = 0{,}1{,}2$ a $n = 0{,}1{,}2$.

Použijeme funkciu *arfima()* z balíčku *arfima*, ktorá počíta odhady metódou maximálnej vierohodnosti. Vyberieme model s najmenšou hodnotou informačných kritérií AIC a BIC. V tabuľke 4.1 sú zobrazené jednotlivé hodnoty pre dané modely. Vyberáme model ARFIMA(0,d,0). Pri odhade pomocou maximálnej vierohodnosti sme použili hustotu z normálneho rozdelenia. Histogram na obrázku

Model	AIC	BIC
m=0, n=0	5640,409	5653,901
m=0, n=1	5641,031	5659,018
m=0, n=2	5642,341	5664,825
m=1, n=0	5641,207	5659,194
m=1, n=1	5642,554	5665,037
m=1, n=2	5643,135	5670,115
m=2, n=0	5641,400	5663,884
m=2, n=1	5642,265	5669,245
m=2, n=2	5637,090	5668,567

Tabuľka 4.1: Informačné kritéria pre rôzne modely z dát z Nílu, balíček *arfima*

4.5, ktorý je v tvare hustoty normálneho rozdelenia, podporuje korektnosť zvolenej metódy.

Procedúra odhadla parameter d na 0,39 a strednú hodnotu na $\hat{\mu} = 1148{,}14$. Dostávame teda model

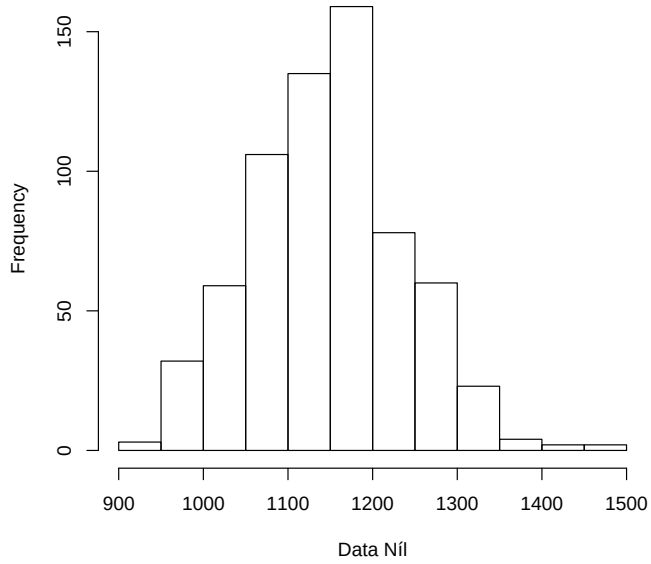
$$(1 - B)^{0.39}(X_t - 1148{,}14) = Y_t,$$

kde Y_t je biely šum s rozptylom 4901,27.

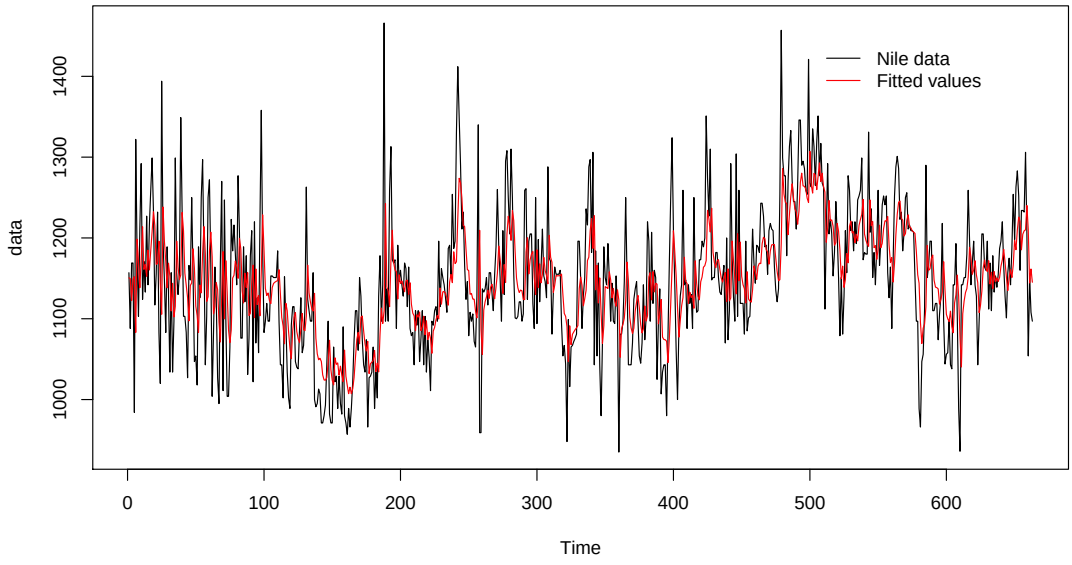
Pôvodný rad aj s odhadnutými hodnotami je zobrazený na obrázku 4.6, pre lepšiu čitateľnosť grafu sme vykreslili aj priebeh prvých 50 hodnôt na obrázku 4.7. Pozorujeme, že odhadnuté hodnoty verne modelujú celkovú dynamiku radu.

Na záver ešte spravme rýchlu diagnostiku modelu na základe vypočítaných reziduí. Z obrázku 4.8 vidíme, že priebeh reziduí skutočne propomína biely šum. Túto skutočnosť podporuje korelogram a graf oneskorených reziduí, na ktorom preloženú priamku takmer nevidieť, splýva s priamkou $x = 0$. Oba grafy vypovedajú o nekorelovanosti reziduí. Túto hypotézu sme potvrdili Boxovým-Ljungovým testom, ktorý na 95% hladine spoľahlivosti nezamietol nulovú hypotézu nekorelovanosti s p-value 0,9. T-testom sme overili predpoklad nulovosti strednej hodnoty, ktorý s p-value 0,76 hypotézu na 95% hladine nezamietol. Konštatný trend z obrázku v pravom dolnom rohu, ktorý znázorňuje druhé mocniny reziduí proti odhadnutným hodnotám (fitted values), pozvrdzuje predpoklad homoskedasticity. Dospeli sme k záveru, že nami odhadnutý model spĺňa všetky predpoklady na reziduálnu zložku.

Ďalšou možnosťou je využitie balíčku *fracdiff*. Obsahuje funkciu *fracdiff()*, ktorá odhaduje parametre takisto MLE metódou, avšak využíva Haslettovu-

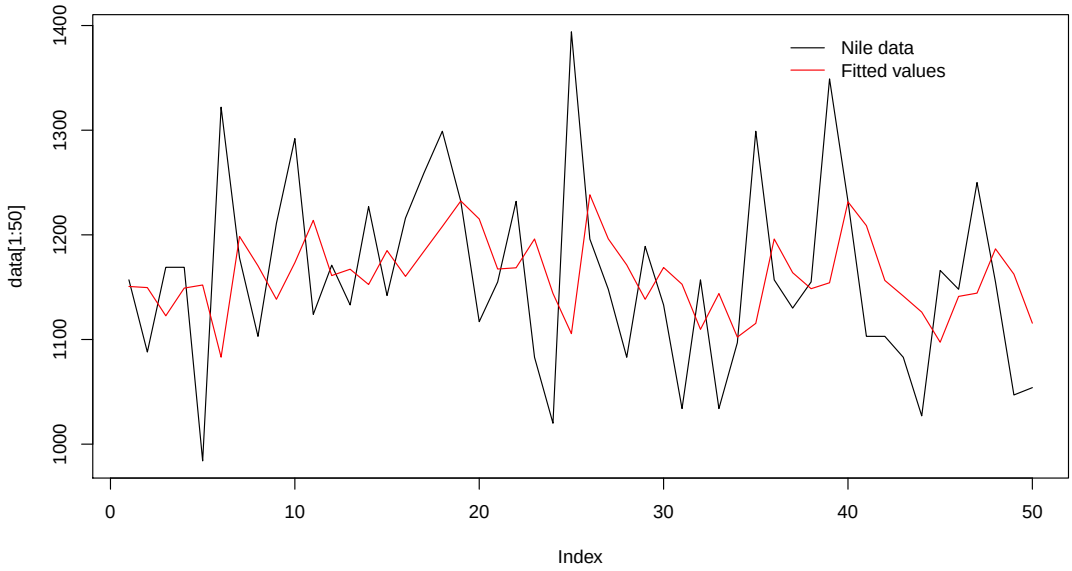


Obr. 4.5: Histogram dát z Nílu



Obr. 4.6: Dáta z Nílu a odhadnuté hodnoty odhadom MLE, model ARFIMA(*0,d,0*), balíček *arfima*

Rafteryho aproximáciu vierohodnostnej funkcie. Takisto ako pri balíčku *arfima* aj tu vytvoríme všetky modely až do rádu dva a následne využijeme funkciu *AIC()* (tento balíček funkciu pre informačné kritérium BIC neobsahuje) na výpočet informačného kritéria. V tabuľke 4.2 sú uvedené jednotlivé hodnoty AIC. Vidíme,



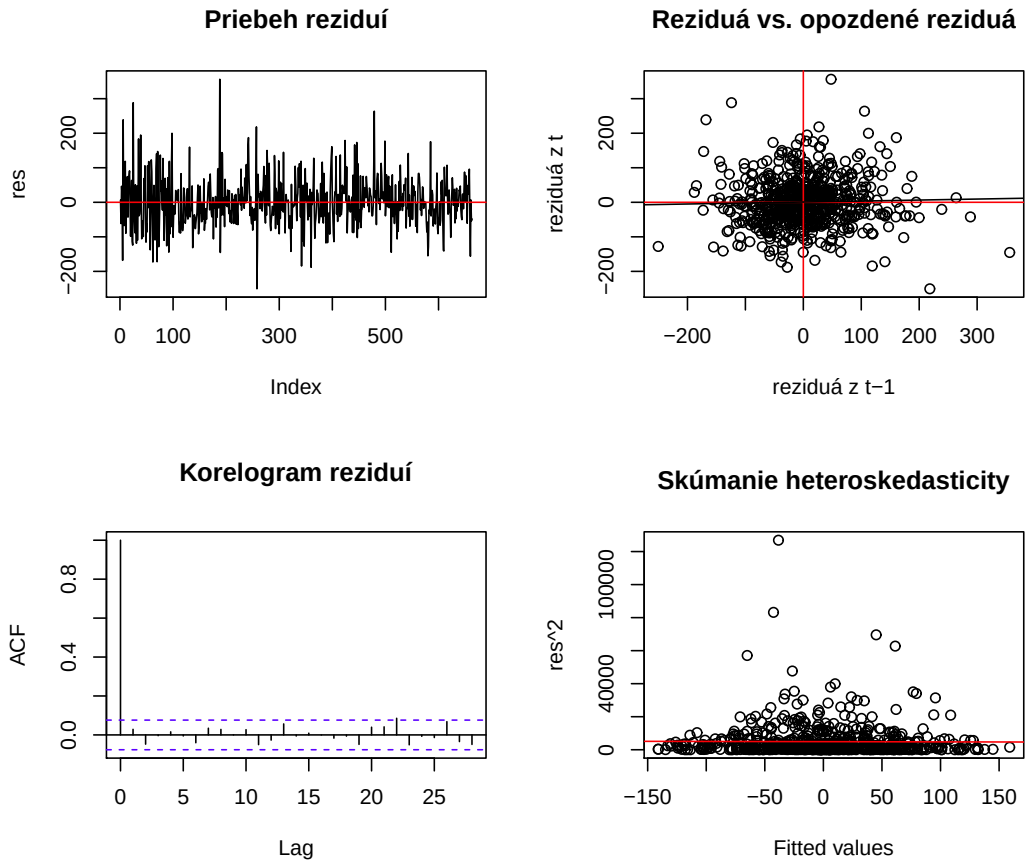
Obr. 4.7: Dáta z Nílu a odhadnuté hodnoty odhadom MLE (prvých 50 hodnôt), model ARFIMA(*0,d,0*), balíček *arfima*

že oproti hodnotám z tabuľky 4.1 sa líšia. Rozdiel spôsobuje rozdielny výpočet vierohodnostnej funkcie. Zároveň poznamenajme, že balíček *arfima* na rozdiel od balíčku *fracdiff* odhaduje aj strednú hodnotu a preto sa pri výpočte AIC líši počet voľných parametrov.

Model	AIC
m=0, n=0	7519.523
m=0, n=1	7522.125
m=0, n=2	7523.672
m=1, n=0	7522.314
m=1, n=1	7523.588
m=1, n=2	7525.64
m=2, n=0	7523.646
m=2, n=1	7525.531
m=2, n=2	7513.998

Tabuľka 4.2: Informačné kritéria pre rôzne modely z dát z Nílu, balíček *fracdiff*

Na základe informačných kritérii volíme model ARFIMA(2,d,2). Parameter d bol odhadnutý na hodnotu 0,38, čo je podobný výsledok ako v minulom prípade.



Obr. 4.8: Reziduálna analýza dát z Nílu, model ARFIMA(*0,d,0*), balíček *arfima*

Autoregresné parametre boli odhadnuté nasledovne: $a_1 = 0{,}83$, $a_2 = 0{,}92$ a parametre kĺzavých súčtov $b_1 = 0{,}87$, $b_2 = 0{,}93$. Máme druhý model s odhadnutou strednou hodnotou 1148,13

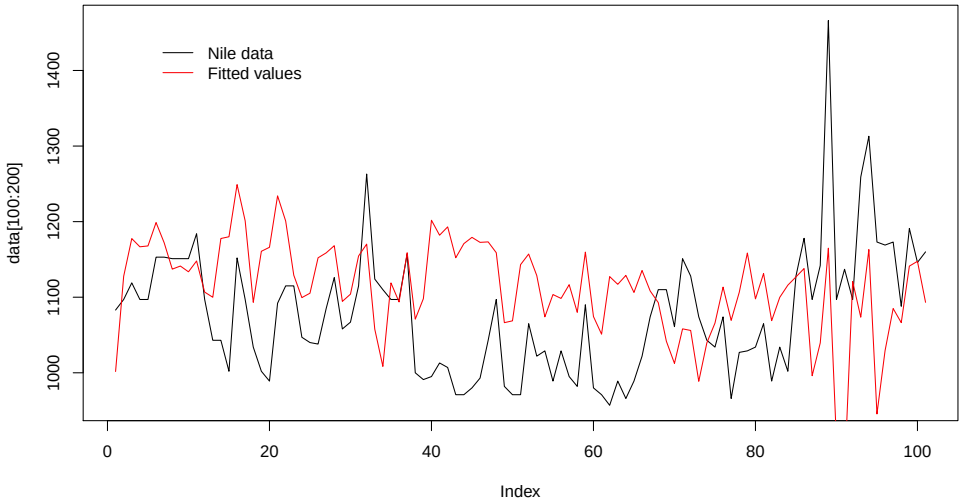
$$(-3157{,}36 + X_t + 0{,}83X_{t-1} + 0{,}92X_{t-2})(1-B)^{0{,}38} = Y_t + 0{,}87Y_{t-1} + 0{,}93Y_{t-2},$$

kde Y_t je biely šum WN(0,σ^2) s vypočítanou smerodatnou odchýlkou $\sigma = 69{,}25$. Odhadnuté hodnoty si vykreslíme v ďalšom balíčku.

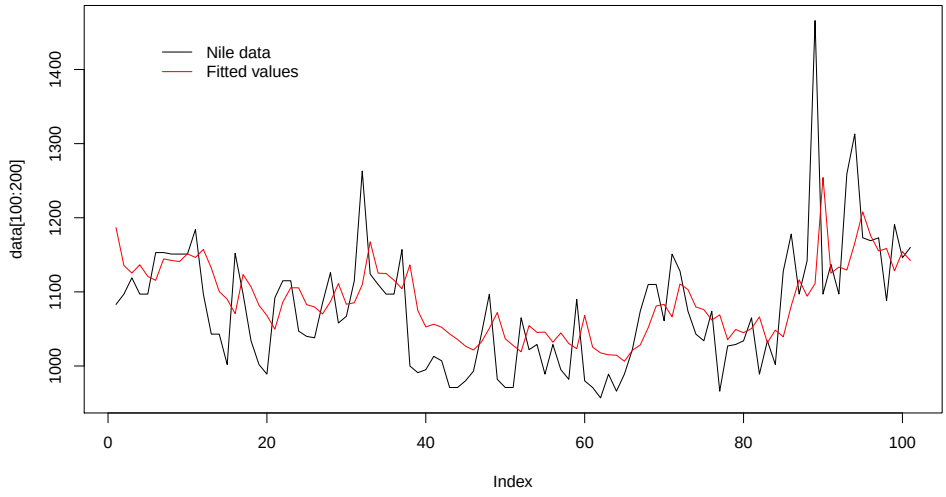
Ďalej uveďme balíček *forecast*. Je v ňom implementovaná procedúra *arfima()*. Po aplikovaní tejto procedúry na dáta z Nílu dostaváme pri voľbe odhadu H-R metódou rovnaký model ako pri balíčku *fracdiff*. Pri voľbe presnou metódou maximálnej vierohodnosti dostávame odhady postupne: $d = 0{,}38$, $a_1 = 0{,}85$, $a_2 = 0{,}91$, $b_1 = 0{,}89$ a $b_2 = 0{,}92$. Máme tretí model s odhadnutou strednou hodnotou 1148,13

$$(-3168{,}84 + X_t + 0{,}85X_{t-1} + 0{,}91X_{t-2})(1-B)^{0{,}38} = Y_t + 0{,}89Y_{t-1} + 0{,}92Y_{t-2},$$

kde smerodatná odchýlka bieleho šumu je 69,25. Na obrázkoch 4.9 a 4.10 sme



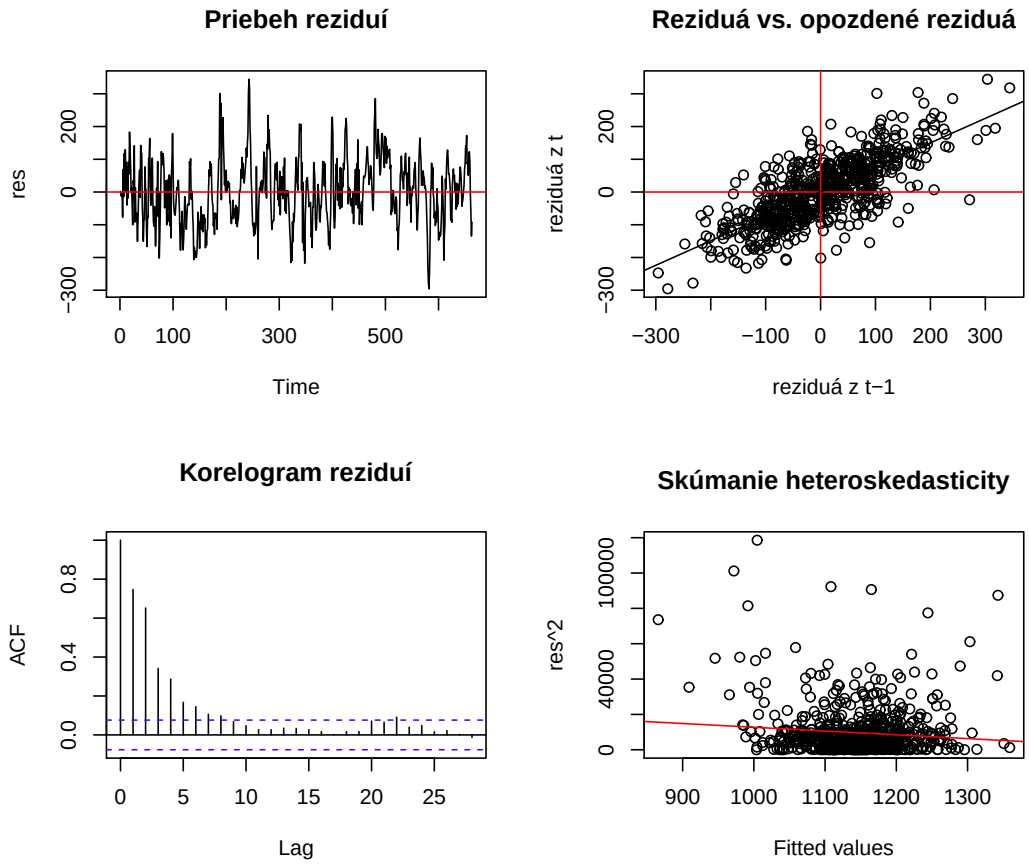
Obr. 4.9: Dáta z Nílu a odhadnuté hodnoty H-R odhadom (výber o rozsahu 100), model ARFIMA(*2,d,2*), balíček *forecast*



Obr. 4.10: Dáta z Nílu a odhadnuté hodnoty presným odhadom MLE (výber o rozsahu 100), model ARFIMA(*2,d,2*), balíček *forecast*

vykreslili modelom odhadnuté hodnoty proti pôvodnej rade na vzorke 100 pozorovaní.

Na obrázku 4.9 pozorujeme, že krátkodobá dynamika nebola Haslettovou-Rafteryho metódou zachytená vhodným spôsobom. Proti tomuto odhadu hovorí aj reziduálna analýza. Na korelograme z obrázku 4.11 jasne vidno korelovanosť reziduí, takisto aj lineárnu závislosť na oneskorených reziduách. Tento fakt po-

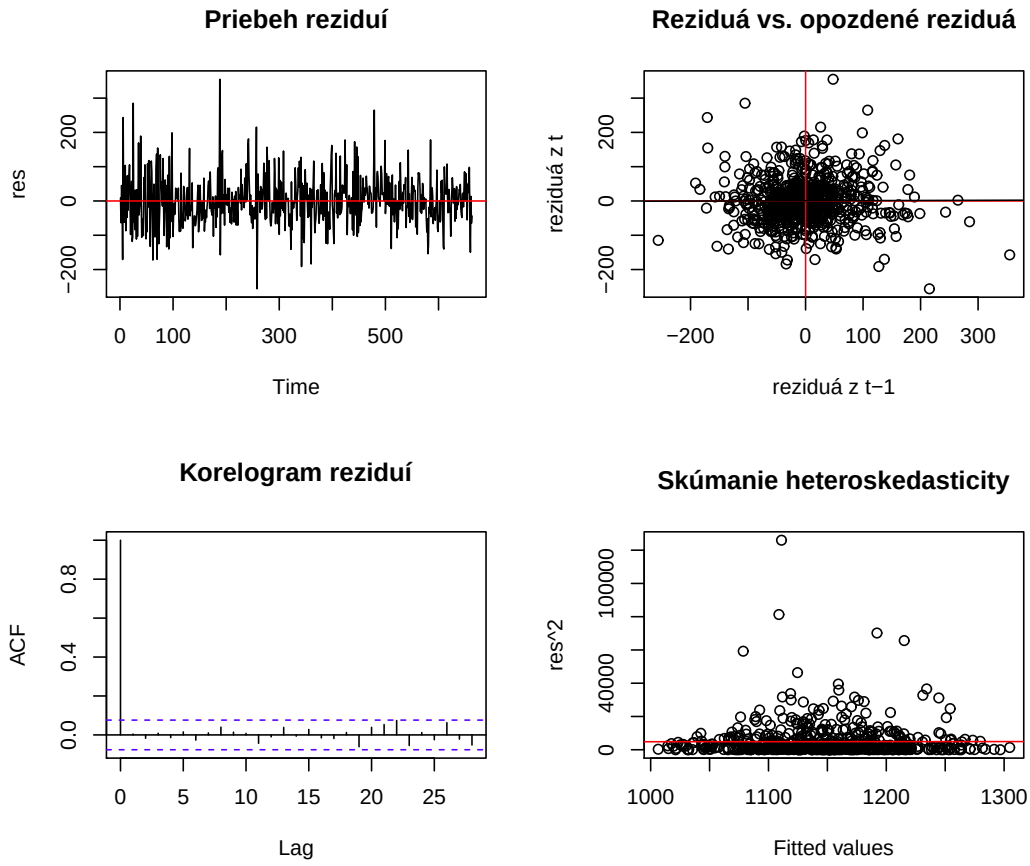


Obr. 4.11: Reziduálna analýza dát z Nílu pre odhad H-R metódou, model ARFIMA(*2,d,2*), balíček *forecast*

tvrdil aj Boxov-Ljungov test, ktorý na hladine spoľahlivosti 95% zamietol hypotézu nekorelovanosti. Z obrázku v pravom dolnom rohu, na ktorom sú vykreslené druhé mocniny reziduí oproti odhadhnutým hodnotám (fitted values), usudzujeme vzhľadom na vzájomnú lineárnu závislosť, že reziduá sú heteroskedastické. Predpoklady na biely šum sú porušené, preto tento model nie je vhodný pre naše dáta.

Na druhej strane odhad modelu ARFIMA(2,d,2) presnou metódou maximálnej vierohodnosti znovu zachytil krátkodobý vývoj radu vierohodne, viď obrázok 4.10. Reziduálna zložka tohto odhadu sa správa ako biely šum. Nekorelovanosť aj homoskedasticitu môžme usúdiť z obrázku 4.12. Nekorelovanosť sme otestovali Boxovým-Ljungovým testom, ktorý nulovú hypotézu na 95% hladine spoľahlivosti nezamietol s p-value 0,98. T-testom sme preverili nulovosť strednej hodoty, ktorú taktiež nezamietame s vysokou p-value 0,82.

Pre úplnosť na záver použime Whittlovu aproximáciu vierohodnostnej funkcie z balíčku *longmemo* pre odhad maximálnou vierohodnosťou. Tento odhad

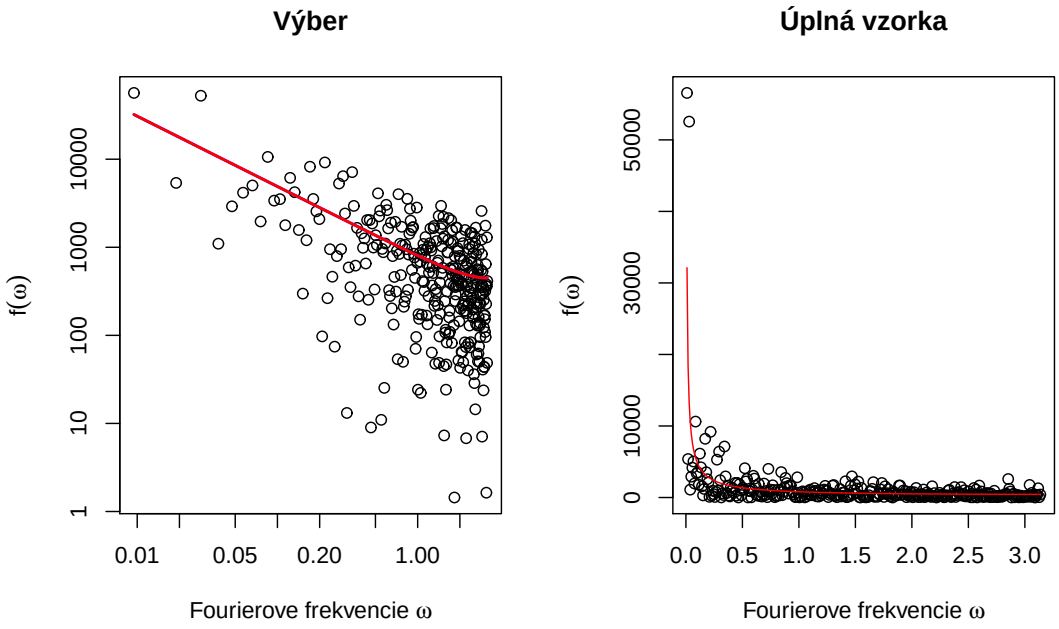


Obr. 4.12: Reziduálna analýza dát z Nílu pre odhad presným MLE, model ARFIMA(*2,d,2*), balíček *forecast*

ukážeme pre názornosť pre model ARFIMA(0,d,0). Po zavolaní funkcie *Whittle-Est* dostávame odhad parametru $H = 0{,}9$ so smerodatnou odchýlkou 0,03. Tento odhad parametru H je takmer totožný s odhadom balíčkom *arfima*. Na obrázku 4.13 sme si nechali vykresliť periodogram a odhad hustoty skonštruovaný z nášho odhadu. Ľavá časť obrázku zobrazuje periodogram len v časti Fourierových frekvencií, pre lepšie zachytenie spektra. V pravej časti je zobrazený periodogram s odhadnutou hustotou pre všetky frekvencie. Z tvaru preloženia priamky je viditeľné, že model dobre zachytil spektrum radu.

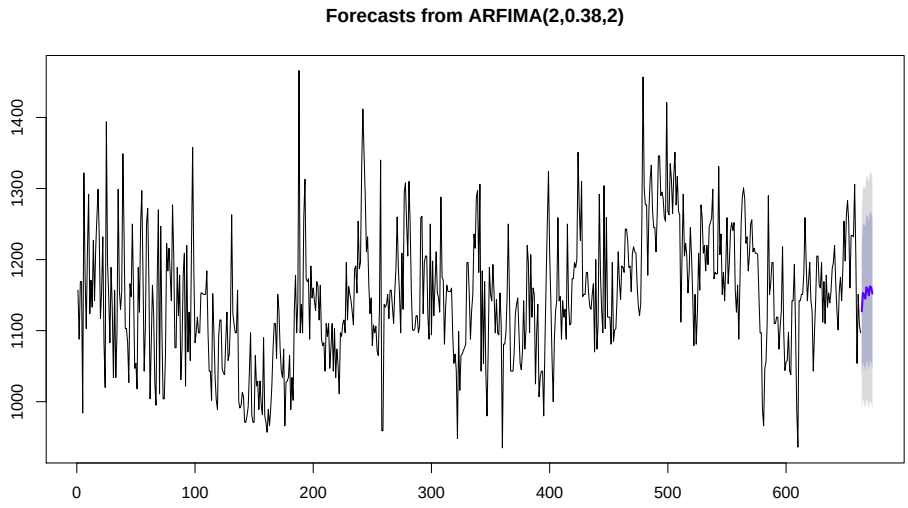
Dospeli sme v skutočnosti k dvom modelom odhadnutým presnou metódou maximálnej vierohodnosti: ARFIMA$(0; 0{,}39; 0)$ a ARFIMA$(2; 0{,}38; 2)$. Parameter d dlhej pamäte bol odhadnutný takmer rovnako. Líši sa len zachytenie krátkodobej dynamiky, t.j. ARMA štruktúry. Z grafov odhadnutých hodnôt vidíme, že model ARFIMA$(2; 0{,}38,; 2)$ lepšie zachytáva skoky, s ktorými si ARFIMA$(0; 0{,}39; 0)$ neporadí.

Balíček *arfima* obsahuje funkciu *predict()* a balíček *forecast* má v sebe im-



Obr. 4.13: Periodogram a odhadnutá spektrálna hustota Whittlovou aproximáciou, model ARFIMA(*0,d,0*), balíček *longmemo*

plementovanú procedúru *forecast()*. Obe procedúry konštruujú ako bodovú tak intervalovú predpoveď. Pre názornú ukážku si ich vykreslíme do grafu predpoveď pomocou funkcie *forecast()*, obrázok 4.14. Tmavomodrá čiara predstavuje bodovú predpoveď, sivá, resp. slabo sivá oblasť je 80%, resp. 95% predikčný interval.



Obr. 4.14: Predpoveď najbližších 10 hodnôt, model ARFIMA(*2,d,2*) balíček *forecast*

Záver

V tejto práci sme si najprv pripravili matematický aparát na definovanie procesu s dlhou pamäťou. Dlhú pamäť sme definovali na základe autokovariančnej funkcie, ukázali sme však aj obecnejší prístup pomocou sebepodobných procesov. Následne sme zaviedli triedu ARFIMA procesov. Pre ARFIMA procesy sme postupne uviedli alebo odvodili základné štatistické charakteristiky a vlastnosti: spektrálnu hustotu, autokovariančnú, resp. autokorelačnú funkciu, prevody na tvar AR(∞) a MA(∞), odhad strednej hodnoty a autokorelačnej funkcie, zaviedli sme predpovede. Ukázali sme, že modelmi ARFIMA môžeme aproximovať iné procesy s dlhou pamäťou.

Ďalej sme predstavili teoretický základ metód odhadov parametrov modelu ARFIMA. Definovali sme známu R/S štatistiku a poukázali na jej nedostatky. Spomenuli sme aj rôzne iné heuristické, grafické metódy detekovania dlhej pamäte. Popri odhade metódou maximálnej vierohodnosti aj s jeho asymptotickými vlastnosťami sme ukázali viaceré možnosti ako aproximovať vierohodnostnú funkciu.

Vo vlastnej práci sme sa venovali analýze dát v štatistickom programe R s hlavným cieľom predstaviť jednotlivé balíčky, ktoré sa zaoberajú procesmi s dlhou pamäťou. Predstavili sme päť vybraných balíčkov a popísali sme ich hlavné funkcie. Ďalej sme použili minimálne ročné hladiny z rieky Níl na predstavenie konkrétnych funkčností. Dospeli sme k viacerým modelom, ktoré sme verifikovali reziduálnou analýzou alebo graficky. Samotný kód z programu R je uložený na priloženom disku.

Zoznam použitej literatúry

[1] Anděl, J. *Základy matematické statistiky*. 2. vydání. Matfyzpress, Praha, 2007.

[2] Beran, J. *Statistic for Long-Memory Processes*. Chapman & Hall, USA, 1994.

[3] Beran, J, Whitcher, B. a Maechler, M. *longmemo: Statistics for Long-Memory Processes (Jan Beran) – Data and Functions*. R package version 1.0-0, http://CRAN.R-project.org/package=longmemo, 2011.

[4] Brockwell, P.J., David, R.A. *Time Series: Theory and Methods*. Colorado State University, Fort Collins, Colorado, 1987.

[5] Cipra, T. *Finační ekonometrie*. Ekopress, s.r.o., Praha, 2008.

[6] Dahlhaus, R. *Efficient paramater estimation for self-similar processes*. Ann. Statist. **17**, 1989.

[7] Diethelm Wuertz and many others and see the SOURCE file. *fArma: ARMA Time Series Modelling*. R package version 3010.79, http://CRAN.R-project.org/package=fArma, 2013.

[8] Fraley, C. Leisch, F., Maechler, M., Reisen V. a Lemonte, A. *fracdiff: Fractionally differenced ARIMA aka ARFIMA(p,d,q) models*. R package version 1.4-2, http://CRAN.R-project.org/package=fracdiff, 2012.

[9] Geweke, J., Porter-Hudak, S. *The stimation and application of long memory time series models*. J. Time Series Anal. **4**, 221-238, 1983.

[10] Gradshteyn, I.S., Ryzhik, I.M. *Tables of Integrals, Series and Products*. Academic Press, San Diego, 2000.

[11] Granger, C.W.J, Joyeux, R. *An introduction to long-range time series models and fractional differencing*. J. Time Series Anal. **1**, 1980.

[12] Grenader, U., Szego, G. *Toeplitz Forms and Their Application.* University of California Press, Berkley, 1958.

[13] Hyndman, R.J., Athanasopoulos, G., Razbash, S., Schmidt,D., Zhou, Z., Khan, Y. a Bergmeir, C. *forecast: Forecasting functions for time series and linear models.* R package version 4.8, http://CRAN.R-project.org/package=forecast, 2013.

[14] Kokoszka, P.S., Taqqu, M.S. *Fractional ARIMA with stable innovations.* Stochastic Process. Appl. **60**, 1995.

[15] Lo, A.W. *Long-term memory in stock market prices.* Econometrica **59**,1279-1313, 1991.

[16] Palma, W. *Long-Memory Time Series.* John Wiley & Sons, Inc., USA, 2007.

[17] Prášková, Z. *Základy náhodných procesů II.* Nakladatelství Karolinum, Praha, 2004.

[18] R Development Core Team: *R: A language and environment for statistical computing.* www.R-project.org, verzia 3.0.2.

[19] Veenstra J. Q. *Persistence and Anti-persistence: Theory and Software.* Western University, http://CRAN.R-project.org/package=arfima, 2012.